Statistical Methods for Joint Analysis of Survival Time and Longitudinal Data

Jaeun Choi

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2011

Approved by:

Advisor: Dr. Jianwen Cai

Advisor: Dr. Donglin Zeng

Reader: Dr. David Couper

Reader: Dr. Bahjat Qaqish

Reader: Dr. Andrew F. Olshan

Abstract

JAEUN CHOI: Statistical Methods for Joint Analysis of Survival Time and Longitudinal Data.
(Under the direction of Dr. Jianwen Cai and Dr. Donglin Zeng.)

In biomedical studies, researchers are often interested in the relationship between patients' characteristics or risk factors and both longitudinal outcomes such as quality of life measured over time and survival time. However, despite the progress in the joint analysis for longitudinal data and survival time, investigation on modeling approach to find which factor or treatment can simultaneously improve the patient's quality of life and reduce the risk of death has been limited. In this dissertation, we investigate joint modeling of longitudinal outcomes and survival time. We consider the generalized linear mixed models for the longitudinal outcomes to incorporate both continuous and categorical data and the stratified multiplicative proportional hazards model for the survival data. We study both Gaussian process and distribution free approaches for the random effect characterizing the joint process of longitudinal data and survival time.

We consider three estimation approaches in this dissertation. First, we consider the maximum likelihood approach with Gaussian process for random effects. The random effects, which are introduced into the simultaneous models to account for dependence between longitudinal outcomes and survival time due to unobserved factors, are assumed to follow a multivariate Gaussian process. The full likelihood, obtained by integrating the complete data likelihood over the random effects, is used for estimation. The Expectation-Maximization (EM) algorithm is used to compute the point

estimates for the model parameters, and the observed information matrix is adopted to estimate their asymptotic variances. Second, the normality assumption of random effects in the likelihood approach is relaxed. Assuming the underlying distribution of random effects to be unknown, we propose using a mixture of Gaussian distributions as an approximation in estimation. Weights of the mixture components are estimated with model parameters using the EM algorithm, and the observed information matrix is used for estimation of the asymptotic variances of the proposed estimators. For the two maximum likelihood approaches with and without normality assumption of random effects, asymptotic properties of the proposed estimators are investigated and their finite sample properties are assessed via simulation studies. Third, we consider a penalized likelihood approach. This approach is expected to be computationally less intensive than the maximum likelihood approach. It gives a penalty for regarding the random effect as a fixed effect in the likelihood and avoids the need to integrate the likelihood over random effects. The penalized likelihood is obtained through Laplace approximation. We compare the numerical performances of the penalized likelihood method and the EM algorithm used in maximum likelihood estimation for the simultaneous models with Gaussian process for random effects via simulation studies. All the proposed methods in this dissertation are illustrated with the real data from the Carolina Head and Neck Cancer Study (CHANCE).

Acknowledgments

I would like to thank my dissertation advisors, Drs. Jianwen Cai and Donglin Zeng, for their expert guidance, deep insights, and thoughtful encouragement throughout my dissertation research process. The lessons I have learned under their direction and the experience I have gained during this process are invaluable. I am grateful to Dr. Jianwen Cai for her financial support as well.

I want to express my sincere gratitude to the committee members, Drs. David Couper, Bahjat Qaqish, and Andrew F. Olshan, for their constructive and perceptive comments. In particular, I am thankful to Dr. Andrew F. Olshan for providing the CHANCE data.

I would like to extend my heartfelt gratitude to the faculties and staffs of the Department of Biostatistics at UNC-CH for their help and encouragement in various ways during my study. I am also grateful to my former colleagues at the UNC Lineberger Comprehensive Cancer Center, where I had my first graduate research assistantship, for their warm consideration and motivation. I want to give my special thanks to all my friends and family for their love, support, and prayer, and to God who has always walked with me through the psalm 119:165, "Great peace have those who love your law, and nothing can make them stumble."

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

The models for jointly analyzing the longitudinal data and survival time have been intensively developed in recent literature. Most models in such analysis would answer the questions regarding how one's quality of life affects time to death or given one's death time how quality of life changes over time. In many biomedical studies, it is of interest to assess the simultaneous effect of treatment or other factors on both patients' quality of life and risk of death taking into account the dependence between quality of life and survival time within a patient. To answer such questions, we consider the simultaneous modeling of longitudinal outcomes and survival time. In this dissertation, we propose two different maximum likelihood approaches with Gaussian process and without distributional assumption for the random effect. In addition, we consider penalized likelihood approach and compare its numerical performance with EM algorithm used in maximum likelihood estimation for the simultaneous models with Gaussian process for the random effect.

1.1 Joint Analysis for Survival Time and Longitudinal Categorical Measurements of Quality of Life in Head and Neck Cancer Patients

Patient survival and Quality of Life (QoL) are often recognized as two major outcome variables in the evaluation of head and neck cancer treatment in oncology community. QoL is important because it reflects the patients' critical physical, psycho–social, and emotional functions and it impacts communication with their caregivers. For the Carolina Head and Neck Cancer Study (CHANCE), we consider a joint analysis of survival time and longitudinal categorical QoL outcomes to find important variables for predicting both patients' QoL and risk of death. We first propose the maximum likelihood approach to simultaneously model the survival time with a stratified Cox proportional hazards model and longitudinal categorical outcomes with a generalized linear mixed model through random effects with normality assumption. Random effects, which are introduced into the simultaneous models to account for dependence between longitudinal outcomes and survival time due to unobserved factors, are assumed to follow a multivariate Gaussian process so that we can use the full likelihood for estimation by integrating the complete data likelihood over the random effects. EM algorithm is used to derive the point estimates for the model parameters, and the observed information matrix is adopted to estimate their asymptotic variances. The asymptotic properties of the proposed estimators are investigated and their finite sample properties are assessed via simulation studies. We illustrate the proposed approach with the real data of longitudinal Head and Neck Cancer Specific symptoms (HNCS) QoL and survival time from the CHANCE study.

1.2 Joint Modeling of Survival Time and Longitudinal Outcomes with Flexible Random Effects

In addition to the maximum likelihood approach with Gaussian process for random effects, we investigate a different method without any distributional assumption for random effects. Gaussian distribution is a convenient distribution often used for the random effects characterizing the joint process of longitudinal outcomes and survival time, and the likelihood approach relies heavily on the such normality assumption. However, this assumption may not be satisfied and the results could be misleading if the assumption is violated. These concerns motivate us to seek more robust estimation method which is not sensitive to the distributional assumption of random effects. Therefore, we relax the normality assumption of random effects by assuming the underlying distribution to be unknown. We propose to use a mixture of Gaussian distributions as an approximation in the estimation. Weights of the mixture components are estimated with model parameters using the EM algorithm, and the observed information matrix is used for the estimation of the asymptotic variances of the proposed estimators. The asymptotic properties of the proposed estimators are investigated and the method is demonstrated to perform well in finite samples via simulation studies. We also conduct simulation studies to examine the robustness of the mixture distribution. AIC and BIC criteria are adopted for selecting the number of mixtures, and the selection procedures are assessed through simulation studies. The proposed method is applied to the CHANCE study aforementioned.

1.3 Penalized Likelihood Approach for Joint Analysis of Survival Time and Longitudinal Outcomes

We compare the numerical performances of the EM algorithm used in the maximum likelihood estimation with Gaussian process for random effect and another estimation method using the penalized likelihood. The penalized likelihood is expected to have less burden on computation because it treats the random effect as the fixed effect in the likelihood and penalized it. Consequently, no calculation is needed to integrate the likelihood over random effects. In SAS GLIMMIX procedure, penalized quasi-likelihood imposing the penalty in quasi-likelihood is already built and used for the analysis of the generalized linear mixed model. Accordingly, it is worthwhile to compare the numerical performances of the EM algorithm and the penalized likelihood method in maximum likelihood estimation. If the EM algorithm performs similarly to the penalized likelihood method on computational time, it will be better to maximize the full likelihood rather than the penalized likelihood. In the meantime, if the penalized likelihood method takes less time and provides unbiased and consistent estimates similar to those from EM algorithm, the penalized likelihood method will be preferred. We present the penalized likelihood obtained through Laplace approximation for our joint models and conduct simulation studies for performance comparison of penalized likelihood and EM algorithm. We also illustrate this comparison through the data analysis of the CHANCE study.

Chapter 2

LITERATURE REVIEW

In this section, we review the statistical literature for : 1) failure time models, 2) longitudinal data models and methods, 3) joint models of failure time and longitudinal data, and 4) penalized quasi-likelihood approach. The organization of the rest of this section is as following. We review literature on statistical methods for Cox proportional hazard models of univariate failure time and frailty models of correlated failure times in section 2.1, and, for generalized linear models with random effects and parameter estimation of longitudinal data in section 2.2. In section 2.3, we review the literature on statistical methods for joint models of failure time and longitudinal data. Lastly, we review penalized quasi-likelihood approach for generalized linear mixed model and bias correction for the estimator in section 2.4.

2.1 Failure Time Models

Failure time analysis or survival analysis addresses data of the form 'time until an event occurs.' The approaches were primarily developed in the medical and biological sciences, but are also broadly used in the social and economic sciences and engineering. A research question arising frequently in these areas is to determine whether or not certain variables are associated with the failure or survival times. There are two major reasons

why this problem cannot be handled via straightforward regression approaches: First, the dependent variable of interest (failure/survival time) is most likely not normally distributed, which is a serious violation of an assumption for ordinary least squares multiple regression. Survival times usually follow a skewed distribution. Second, there is the problem of censoring, that is, some observations will be incomplete.

We summarize the Cox proportional hazard model for the univariate failure time, which is not based on any assumptions concerning the nature or shape of the underlying survival distribution, in section 2.1.1, and the frailty model for the correlated failure time, which formulates the nature of dependence explicitly, as an extension of the Cox model in section 2.1.2.

2.1.1 Univariate failure time model

The Cox proportional hazards model (Cox, 1972) has been the most widely used procedure to study the effects of covariates on a failure time. The Cox model assumes that the hazard function for the failure time T associated with a covariate vector $\boldsymbol{Z}$ is given by

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp\{\boldsymbol{\beta}_0^T\boldsymbol{Z}(t)\}, \qquad t \geq 0 \tag{2.1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}_0$ is a $p\times 1$ vector of unknown regression parameters. The model (2.1) is semi-parametric in that the effect of the covariates on the hazard is explicitly specified while the form of the baseline hazard function is unspecified. The model (2.1) assumes that hazard ratios are proportional across groups or subpopulations over time, and the regression coefficient $\boldsymbol{\beta}_0$ represents the log hazard ratio for one unit increase in the corresponding covariate given that the other covariates in the model are held at the same value.

Let C denote the potential censoring time and $X = \min(T, C)$ denote the observed time. Let $N(t)$ denote the counting process, $Y(t) = I(X \leq t)$ be an 'at-risk' indicator

process and $\Delta = I(T \leq C)$ be an indicator for failure, where $I(.)$ is an indicator function. The failure time is assumed to be subject to independent right censorship. Let $(T_i, C_i, \boldsymbol{Z}_i)(i = 1, \ldots, n)$ be n independent replicates of $(T, C, \boldsymbol{Z})$ and τ denote the study end time.

The regression parameter $\boldsymbol{\beta}_0$ in (2.1) can be estimated by applying standard asymptotic likelihood procedure to the 'partial' likelihood function, introduced by Cox (1975),

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp\{\boldsymbol{\beta}^T \boldsymbol{Z}_i(T_i)\}}{\sum_{l=1}^{n} Y_l(T_i) \exp\{\boldsymbol{\beta}^T \boldsymbol{Z}_l(T_i)\}} \right]^{\Delta_i},$$

where $\boldsymbol{Z}_i(T_i)$ is the covariate vector for the subject failing at T_i, and $\boldsymbol{Z}_l(T_i)$ is the corresponding covariate vector for the l-th member who is at risk at T_i. The estimator for $\boldsymbol{\beta}_0$, denoted by $\hat{\boldsymbol{\beta}}$, is obtained by the partial likelihood score function

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \left\{ \boldsymbol{Z}_i(X_i) - \frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, X_i)}{S^{(0)}(\boldsymbol{\beta}, X_i)} \right\},$$

where $S^{(0)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) \exp\{\boldsymbol{\beta}' \boldsymbol{Z}_i(t)\}$, $\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} Y_i(t) \exp\{\boldsymbol{\beta}' \boldsymbol{Z}_i(t)\}$ $\boldsymbol{Z}_i(t)$. The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$, defined as the solution to the unbiased score equation $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$, has been shown to be approximately normal in large samples with mean $\boldsymbol{\beta}_0$ and with a covariance matrix that can be consistently estimated by $-\left\{ \frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} \right\}^{-1}$ (Andersen & Gill, 1982; Tsiatis, 1981). Iterative procedures, such as Newton-Raphson method and EM algorithm, are commonly used to solve the score equation.

Cox proportional hazards model has been extended from analyzing univariate failure time data to multivariate failure time data. Andersen & Gill (1982) and Fleming & Harrington (1991) extended Cox model in the expression of counting process which is more general and includes recurrent failures. Wei, Lin & Weissfeld (1989) and Hougaard (2000) extended Cox model to model multivariate failure times. As another extension

from Cox model, Clayton & Cuzick (1985) and Hougaard (2000) proposed frailty model for clustered failure data in which subjects may or may not experience the same type of event but they may be correlated because subjects are from the same cluster. In the frailty model, Cox proportional hazards model is used to model each individual's hazard function, and then an unobserved cluster-specific frailty is introduced into each model to account for within-cluster correlation. This frailty model is reviewed in the next section 2.1.2.

2.1.2 Correlated failure time model

The Cox model (2.1) in the previous section 2.1.1 assumes the independent failure times. In many biomedical studies, however, the independence between failure times might be violated, which may arise because study subjects may be grouped in a manner that leads to dependencies within groups, or because individuals may experience multiple events. For such data, there are two main approaches: the marginal model approach which leaves the nature of dependence among related failure times completely unspecified and the frailty model approach which formulate the nature of dependence explicitly. When the interest resides in estimating the effect of risk factors and the correlation among the failure times are considered as a nuisance, the marginal model approach suits this purpose very well. However, in some settings, one might be interested in the strength and nature of dependencies among the failure time components, for which the frailty models have been proposed and studied by many authors. We focus on frailty model in this section.

The frailty model explicitly formulates the nature of the underlying dependence structure through an unobservable random variable. This unknown factor is usually called individual heterogeneity or frailty. The key assumption is that the failure times are conditionally independent given the value of the frailty. To illustrate this idea,

consider a Cox proportional hazards model for subject i with respect to the kth event :

$$\lambda_{ik}(t|w_i) = w_i \lambda_0(t) \exp\{\boldsymbol{\beta}_0^T \boldsymbol{z}_{ik}(t)\} \tag{2.2}$$

where the frailty terms $\{w_i\}, i = 1, \ldots, n$ are assumed to be independent and to arise from a common parametric density. The commonly used one is the gamma distribution, mostly for mathematical convenience. Various choices are possible for this density, which include the positive stable distributions, the inverse Gaussian distributions and the log-normal distributions. Note that $\boldsymbol{\beta}_0$ in (2.2) generally needs to be interpreted conditionally on the unobserved frailty. The frailty model approach is particularly sensible, when the strength of the dependence of failure times is of interest.

The parameter estimates are obtained through the EM algorithm, making use of the partial likelihood expression in the maximization step as shown in Klein (1992). An alternative approach is to use a penalized partial likelihood for the estimation of the shared frailty (Therneau & Grambsch, 2001).

Troxel & Esserman (2004) proposed a novel application of frailty models to assess the correlation between survival and quality of life in oncology. A frailty parameter is a random effect that allows the variability among clusters of measurements to be incorporated into survival models. The collected quality of life outcomes are dichotomized in order to apply the multivariate survival methods. In spite of the necessity of the conversion, the discretization of the quality of life scores from a continuous to a failure-time structure leads to the loss of information available from continuous quality of life data.

Ratcliffe, Guo, & TenHave (2004) proposed a joint model for the analysis of longitudinal and survival data in the presence of data clustering via a common frailty. While the existing models include subject-level random effects as the only random effects, two levels of nested random effects (subject-level random effects and cluster-level

frailty, with subjects nested within clusters) are used in the model with the responses linked at the higher cluster level. This additional level of random effects makes the model more flexible. They used a mixed effects model for the repeated measures that incorporates both subject- and cluster-level random effects, with subjects nested within clusters. A Cox frailty model is used for the survival model as it allows for between-cluster heterogeneity. Then they link the two responses via the common cluster-level random effects, or frailties, using a multivariate normal assumption for computation ease (Li & Lin, 2000). More joint models of survival and longitudinal data are reviewed using different models in section 2.3.

2.2 Longitudinal Data Models and Methods

The defining feature of a longitudinal study is that individuals are measured repeatedly through time. Longitudinal data require special statistical methods because the set of observations on one subject tends to be intercorrelated. This correlation must be taken into account to draw valid scientific inferences.

There are a variety of qualitatively different sources of random variation that might actually occur in practice and be included to construct the model for longitudinal data. The linear random effects model described in section 2.2.1 is one of three extensions of generalized linear models for longitudinal data: marginal, random effects, and transition models. The random effects model is most useful when the objective is to make inference about individuals rather than the population average. The parameter estimation in the generalized linear model with random effects can be carried out by both maximum likelihood and conditional likelihood approaches summarized in section 2.2.2.

2.2.1 Generalized linear model with random effects

The linear random effects model is applied where the response is assumed to be a linear function of explanatory variables with regression coefficients that vary from one individual to the next. This variability reflects natural heterogeneity due to unmeasured factors, which can be represented by a probability distribution. Correlation among observations for one person arises from their sharing unobservable variables, U_i.

The random effects GLM has the following general specifications:

1. Given the random effects $\boldsymbol{U}_i$, the responses $Y_{i1}, \ldots, Y_{in_i}$ are mutually independent and follows a distribution from the exponential family with density

$$f(y_{ij}|\boldsymbol{U}_i;\boldsymbol{\beta}) = \exp[\{(y_{ij}\theta_{ij} - \psi(\theta_{ij}))\}/\phi + c(y_{ij},\phi)]. \tag{2.3}$$

The conditional moments, $\mu_{ij} = \mathrm{E}(Y_{ij}|\boldsymbol{U}_i) = \psi'(\theta_{ij})$ and $v_{ij} = \mathrm{Var}(Y_{ij}|\boldsymbol{U}_i) = \psi''(\theta_{ij})\phi$, satisfy $h(\mu_{ij}) = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{d}_{ij}^T\boldsymbol{U}_i$ and $v_{ij} = v(\mu_{ij})\phi$ where h and v are known link and variance functions, respectively, and $\boldsymbol{x}_{ij}$ and $\boldsymbol{d}_{ij}$ are covariate vectors of length p and q, respectively. $\boldsymbol{d}_{ij}$ is a subset of $\boldsymbol{x}_{ij}$.

2. The random effects, $\boldsymbol{U}_i$, $i = 1, \ldots, m$, are mutually independent and identically distributed with density function $f(\boldsymbol{U}_i; G)$.

Another fundamental assumption of the random effects model is that the $\boldsymbol{U}_i$ are independent of the explanatory variables. A model of this type is sometimes referred to as a "latent variable" model (Bartholomew, 1987).

2.2.2 Maximum likelihood and conditional likelihood methods

Let $\boldsymbol{U} = (U_1, ..., U_m)$. In maximum likelihood approach, $\boldsymbol{U}$ is treated as a set of unobserved variables which is integrated out of the likelihood, adopting the assumption that the random effects follow a distribution such as Gaussian model with mean zero and

variance matrix $\boldsymbol{G}$. In conditional likelihood approach, the random effects is treated as if they were fixed parameters to be removed from the problem, so that we need not rely on the second assumption in the previous section 2.2.1.

Maximum likelihood approach treats $\boldsymbol{U}_i$ as a sample of independent unobservable variables from a random effects distribution. Then, the likelihood function for the unknown parameter $\boldsymbol{\delta}$, which is defined to include both $\boldsymbol{\beta}$ and the elements of $\boldsymbol{G}$, is

$$L(\boldsymbol{\delta};\boldsymbol{y}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{U}_i;\boldsymbol{\beta}) f(\boldsymbol{U}_i;\boldsymbol{G}) \mathrm{d}\boldsymbol{U}_i, \tag{2.4}$$

which is the marginal distribution of $\boldsymbol{Y}$ obtained by integrating the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{U}$ with respect to $\boldsymbol{U}$. In some special case such as the Gaussian linear model, the integral in (2.4) has a closed form, but for most non-Gaussian models, numerical methods are required for its evaluation.

To find the maximum likelihood estimate, we solve the score equations obtained by setting to zero the derivative with respect to $\boldsymbol{\delta}$ of the log likelihood. Considering the 'complete' data for an individual to comprise $(\boldsymbol{y}_i, \boldsymbol{U}_i)$ and restricting attention to canonical link functions (McCullagh & Nelder, 1989) for which $\theta_{ij} = h(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{d}_{ij}^T\boldsymbol{U}_i$, then the 'complete data' score function for $\boldsymbol{\beta}$ has a particularly simple form

$$\boldsymbol{S}_\beta(\boldsymbol{\delta}|\boldsymbol{y},\boldsymbol{U}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \boldsymbol{x}_{ij}\{y_{ij} - \mu_{ij}(\boldsymbol{U}_i)\} = 0, \tag{2.5}$$

where $\mu_{ij}(\boldsymbol{U}_i) = \mathrm{E}(y_{ij}|\boldsymbol{U}_i) = h^{-1}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{d}_{ij}^T\boldsymbol{U}_i)$. The observed data score functions $\boldsymbol{S}_\beta(\boldsymbol{\delta}|\boldsymbol{y})$ are defined as the expectations of the complete data score functions $\boldsymbol{S}_\beta(\boldsymbol{\delta}|\boldsymbol{y},\boldsymbol{U})$ in (2.5) with respect to the conditional distribution of $\boldsymbol{U}$ given $\boldsymbol{y}$. This gives,

$$\boldsymbol{S}_\beta(\boldsymbol{\delta}|\boldsymbol{y}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \boldsymbol{x}_{ij}[y_{ij} - \mathrm{E}\{\mu_{ij}(\boldsymbol{U}_i)|\boldsymbol{y}_i\}] = 0. \tag{2.6}$$

The score equations for $\boldsymbol{G}$ can similarly be obtained as

$$\boldsymbol{S}_G(\boldsymbol{\delta}|\boldsymbol{y}) = \frac{1}{2}\boldsymbol{G}^{-1}\left\{\sum_{i=1}^{m}\mathrm{E}(\boldsymbol{U}_i\boldsymbol{U}_i^T|\boldsymbol{y}_i)\right\}\boldsymbol{G}^{-1} - \frac{m}{2}\boldsymbol{G}^{-1} = 0. \tag{2.7}$$

A common strategy to solve for the maximum likelihood estimate of $\boldsymbol{\delta}$ is to use the EM algorithm (Dempster *et al.*, 1977). This algorithm iterates between an E-step, which involves evaluating the expectations in the above score equations (2.6) and (2.7) using the current values of the parameters, and an M-step, in which we solve the score equations to give updated parameter estimates. The dimension of the integration involved in the conditional expectation is q, the dimension of $\boldsymbol{U}_i$. When q is one or two, numerical integration techniques can be implemented reasonably easily. (e.g. Crouch & Spiegelman, 1990) For higher dimensional problems, Monte Carlo integration methods can be used. (e.g. the application of Gibbs sampling in Zeger & Karim, 1991)

Gaussian distribution is a convenient model used most for the random effects. When the regression coefficients are of primary interest, the specific form of the random effects distribution is less important. However, when the random effects are themselves the focus, inferences are more dependent on the assumptions about their distribution. Lange & Ryan (1989) suggested a graphical way to test the Gaussian assumption when the response variables are continuous. When the response variables are discrete, the same task becomes more difficult. Davidian & Gallant (1992) developed a non-parametric approach to estimate the random effects distribution with non-linear models.

In conditional likelihood approach for the generalized linear models with random effects (Diggle *et al.*, 1994; McCullagh & Nelder, 1989), the main idea is to treat the random effects $\boldsymbol{U}_i$ as a set of nuisance parameters to be removed, and to estimate $\boldsymbol{\beta}$ using the conditional likelihood of the data given the sufficient statistics for the $\boldsymbol{U}_i$.

Treating $\boldsymbol{U}$ as fixed, the likelihood function for $\boldsymbol{\beta}$ and $\boldsymbol{U}$ is

$$L(\boldsymbol{\beta}, \boldsymbol{U}; \boldsymbol{y}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{U}_i) \propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} \exp\{\theta_{ij} y_{ij} - \psi(\theta_{ij})\}, \tag{2.8}$$

where $\theta_{ij} = \theta_{ij}(\boldsymbol{\beta}, \boldsymbol{U})$. Restrict attention to canonical link functions (McCullagh & Nelder, 1989) for which $\theta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{d}_{ij}^T \boldsymbol{U}_i$, the likelihood in (2.8) can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{U}; \boldsymbol{y}) = \exp\left\{ \boldsymbol{\beta}^T \sum_{i,j} \boldsymbol{x}_{ij} y_{ij} + \sum_i \boldsymbol{U}_i^T \sum_j \boldsymbol{d}_{ij} y_{ij} - \sum_{i,j} \psi(\theta_{ij}) \right\}.$$

Hence, the sufficient statistics for $\boldsymbol{\beta}$ and $\boldsymbol{U}_i$ are $\sum_{i,j} \boldsymbol{x}_{ij} y_{ij}$ and $\sum_{i,j} \boldsymbol{d}_{ij} y_{ij}$ respectively, and $\sum_{i,j} \boldsymbol{d}_{ij} y_{ij}$ is sufficient for $\boldsymbol{U}_i$ for fixed $\boldsymbol{\beta}$.

The conditional likelihood is proportional to the conditional distribution of the data given the sufficient statistics for the $\boldsymbol{U}_i$, and the contribution from subject i has the form

$$\begin{aligned} f\left(\boldsymbol{y}_i | \sum_j \boldsymbol{d}_{ij} y_{ij} = \boldsymbol{b}_i; \boldsymbol{\beta} \right) &= \frac{f(\boldsymbol{y}_i; \boldsymbol{\beta}, \boldsymbol{U}_i)}{f(\sum_j \boldsymbol{d}_{ij} y_{ij} = \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{U}_i)} \\ &= \frac{f(\sum_j \boldsymbol{x}_{ij} y_{ij} = \boldsymbol{a}_i, \sum_j \boldsymbol{d}_{ij} y_{ij} = \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{U}_i)}{f(\sum_j \boldsymbol{d}_{ij} y_{ij} = \boldsymbol{b}_i; \boldsymbol{\beta}, \boldsymbol{U}_i)}. \end{aligned} \tag{2.9}$$

For a discrete generalized linear model, this expression (2.9) can be written as

$$P\left(\boldsymbol{y}_i | \sum_j \boldsymbol{d}_{ij} y_{ij} = \boldsymbol{b}_i; \boldsymbol{\beta} \right) = \frac{\sum_{R_{i1}} \exp(\boldsymbol{\beta}^T \boldsymbol{a}_i + \boldsymbol{U}_i^T \boldsymbol{b}_i)}{\sum_{R_{i2}} \exp(\boldsymbol{\beta}^T \sum_j \boldsymbol{x}_{ij} y_{ij} + \boldsymbol{U}_i^T \boldsymbol{b}_i)},$$

where R_{i1} is the set of possible values for $\boldsymbol{y}_i$ such that $\sum_j \boldsymbol{x}_{ij} y_{ij} = \boldsymbol{a}_i$ and $\sum_j \boldsymbol{d}_{ij} y_{ij} = \boldsymbol{b}_i$, and R_{i2} is the set of values for $\boldsymbol{y}_i$ such that $\sum_j \boldsymbol{d}_{ij} y_{ij} = \boldsymbol{b}_i$. The conditional likelihood for $\boldsymbol{\beta}$ given the data for all m individuals simplifies to

$$L(\boldsymbol{\beta} | \boldsymbol{y}, \sum_j \boldsymbol{d}_{ij} y_{ij} = \boldsymbol{b}_i) = \prod_{i=1}^{m} \frac{\sum_{R_{i1}} \exp(\boldsymbol{\beta}^T \boldsymbol{a}_i)}{\sum_{R_{i2}} \exp(\boldsymbol{\beta}^T \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} y_{ij})}. \tag{2.10}$$

For simple cases such as the random intercept model, the conditional likelihood is reasonably easy to maximize (Breslow and Day, 1980). By the analogy with the usual score equations derived from the full likelihood, the score equations obtained from the conditional likelihood (2.10) can be used to get maximum conditional likelihood estimator for $\boldsymbol{\beta}$.

The random effects generalized linear models in biostatistics have been studied enormously including the following literatures providing useful additional references: Laird & Ware (1982); Stiratelli *et al.*(1984); Gilmour *et al.*(1985); Schall (1990); Zeger & Karim (1991); Waclawiw & Liang (1993); Solomon & Cox (1992); Breslow & Clayton (1993); Drum & McCullagh (1993); Breslow & Lin (1995) and Lin & Breslow (1996).

2.3 Joint Models of Failure Time and Longitudinal Data

Joint analysis of survival time and repeated measurements has been intensively studied in recent literature. The most models which have been used in such analysis can be categorized into a selection model or a pattern mixture model. The selection model would answer the question regarding how one's quality of life affects death and the pattern-mixtrure model would describe the pattern of quality of life given one's death time. However, research interest is also often in finding which factor or treatment can simultaneously improve the patients' quality of life and reduce the risk of death, which can be studied by the simultaneous analysis of quality of life and survival.

Let $\boldsymbol{Y}$ denote the longitudinal outcomes, for example, quality of life, then $\boldsymbol{Y}$ are realizations of a latent process $\tilde{\boldsymbol{Y}}$ measured with errors. Let T denote survival time.

A selection model focuses on estimating the distribution of T given $\tilde{\boldsymbol{Y}}$. Such a selection model has been studied by many authors: Tsiatis *et al.*(1995), Wulfsohn and Tsiatis (1997), Hu *et al.*(1998), Huang *et al.*(2001), and Xu and Zeger (2001a, b). Usually, $\tilde{\boldsymbol{Y}}$ is modeled as a function of observed covariates and subject-specific random

effects; then it is fed into the model of T given $\tilde{\boldsymbol{Y}}$ as a linear predictor. Selection model is reviewed in section 2.3.1.

In the pattern mixture model, a model is assumed for longitudinal outcome $\boldsymbol{Y}$ conditional on survival time T (Wu and Carroll, 1988; Wu and Bailey, 1989; Hogan and Laird, 1997) and interest focuses on estimating parameters in the model for longitudinal outcome.

Simultaneous modeling serves the purpose to model both the process for quality of life, $\boldsymbol{Y}$, and survival time, T, given observed covariates $\boldsymbol{X}$. Zeng & Cai (2005) proposed such a model of quality of life $\boldsymbol{Y}$ following normal distribution and survival time T by the observed covariates $\boldsymbol{X}$ and by unobserved factors with normal density. This approach is reviewed in section 2.3.2. It is noted that this approach is different from either selection model or pattern-mixture model, although mathematically, all three models can be regarded as different ways of writing the distribution of (T, $\boldsymbol{Y}$) given covariates.

2.3.1 Failure time model with longitudinal covariates

Many longitudinal studies collect information on each participant both on a time-to-event and covariates which may vary with time. Recent interest has focused on joint models for longitudinal covariate data and a survival endpoint. A popular approach assumes that the longitudinal data follow a linear mixed effects model (Laird & Ware, 1982) and that survival depends on the covariate through a proportional hazards relationship with the underlying random effects. To implement the Cox model with time-dependent covariates, complete knowledge of the true covariate history for each subject is required; however, time-dependent covariates are generally measured intermittently, often at different times for each subject and with error. A naive approach is to substitute for each subject at each failure time in the Cox partial likelihood (Cox,

1975) the closest observed covariate value prior to that time, often termed 'last value carried forward'. It is well known (Prentice, 1982) that substituting mis-measured values for true covaiates in the Cox model leads to biased estimation. Another strategy for estimation of the proportional hazard regression parameters is a two-stage approach (Pawitan & Self, 1993; Tsiatis *et al.*, 1995) : First, the mixed effects model is fitted to data at each risk set assuming normality for both random effects and intra-subject error from which empirical Bayes estimates of the individual random effects are obtained as described by Laird and Ware (1982). Then, predictors for the covariate for each subject at each failure time based on the relevant fit are substituted for the true covariate values in the Cox partial likelihood. This approximate method uses regression calibration (Carroll *et al.*, 1995) to reduce bias of the naive approach but still yields biased estimators for large measurement error. Alternatively, the joint likelihood of the survival and longtidinal data may form the basis for inference. DeGruttola & Tu (1994) assumed the covariate process and survival times to be multivariate normal and fitted the model via parametric maximum likelihood. Wulfsohn & Tsiatis (1997) adopted the less rigid proportional hazards relationship and used nonparametric maximum likelihood, but continued to assumed normal random effects. Henderson *et al.*(2000) used normal random effects in Gaussian covariate stochastic processes. Faucett & Thomas (1996) assumed normality and took a Bayesian approach.

These strategies rely heavily on the assumption of normality of random effects characterizing the true covariate process; however, this assumption may be over-restrictive and the consequences if it is violated are unknown. Tsiatis & Davidian (2001) proposed a conditional score estimation for the proportional hazards model with longitudinal covariates with measurement-errors, which does not put any restrictions on the distribution of the random effects in covariate process by exploiting the conditional score approach of Stefanski & Carroll (1987).

In this section, we focus on the conditional score estimation approach by Tsiatis & Davidian (2001) since the fundamental idea of the maximum likelihood approach, which has been mostly studied with distributional assumption of random effects by many authors, is same as that reviewed in section 2.3.2.

For each subject i $(i = 1, ..., n)$, let T_i and C_i denote time to failure and censoring, respectively, where time on study $V_i = \min(T_i, C_i)$ and failure indicator $\Delta_i = I(T_i \leq C_i)$ are observed; all variables are independent across i. Let Z_i denote time-independent covariates and $X_i(u)$ denote time-dependent covariates at time u for subject i; for simplicity, assume $X_i(u)$ is scalar, but generalization to vector-valued $X_i(u)$ is straightforward. Assume that $X_i(u)$ follows a subject-specific linear model $X_i(u) = \alpha_{0i} + \alpha_{1i}u$, where $\alpha_i = (\alpha_{0i}, \alpha_{1i})^T$ are the intercept and slope for i. The covariate process $X_i(u)$ is not directly observed; rather, longitudinal measurements $W_i(t_{ij})$ are obtained at ordered times $t_i = (t_{i1}, ..., t_{im_i})^T$, for $t_{im_i} \leq V_i$, where $W_i(t_{ij}) = X_i(t_{ij}) + e_{ij}$, with $e_i = (e_{i1}, ..., e_{im_i})^T$. The errors e_{ij} reflect uncertainty in measuring $X_i(u)$ at t_{ij} and are assumed identically normally distributed and independent with mean zero and variance σ^2, independent of $(T_i, C_i, \alpha_i, Z_i, t_i, m_i)$. More precisely,

$$(e_i | T_i, C_i, \alpha_i, Z_i, t_i, m_i) \sim N_{m_i}(0, \sigma^2 I_{m_i}),$$

where $I_{m_i}$the m_i-dimensional identity matrix.

The survival model assumes that the hazard of failure is related to $X_i(u)$ and Z_i through a proportional hazards regression model; that is,

$$\begin{aligned} \lambda_i(u) &= \lim_{du \to 0} du^{-1} \text{pr}\{u \leq T_i < u + du | T_i \geq u, \alpha_i, Z_i, C_i, e_i(u), t_i(u)\} \\ &= \lim_{du \to 0} du^{-1} \text{pr}\{u \leq T_i < u + du | T_i \geq u, \alpha_i, Z_i\} \\ &= \lambda_0(u) \exp\{\gamma X_i(u) + \eta^T Z_i\}, \end{aligned} \tag{2.11}$$

where $\lambda_0(u)$ denotes an unspecified baseline hazard function, the collection of times of longitudinal measurements up to and including u is denoted by $t_i(u) = (t_{ij} \leq u)$, $e_i(u) = (e_{ij} : t_{ij} \leq u)$, and η is $(q \times 1)$. The model (2.11) shows explicitly the nature of the assumption that timing of measurements and censoring are noninformative. Interest focuses on estimation of the parameters γ and η.

Let $\hat{X}_i(u)$ be the ordinary least squares estimator of $X_i(u)$ using all the longitudinal data up to and including time u, that is based on $t_i(u)$. This requires at least two longitudinal measurements on i up to and including u, for $t_{i2} \leq u$. Define the counting process increment

$$dN_i(u) = I(u \leq V_i < u + du, \Delta_i = 1, t_{i2} \leq u)$$

and the 'at risk' process $Y_i(u) = I(V_i \geq u, t_{i2} \leq u)$; that is, $dN_i(u)$ puts point mass at time u corresponding to the observed death time for the i-th subject as long as this occurs after the second longitudinal measurement, and $Y_i(u)$ is the indicator that subject i is at risk with at least two longitudinal measurement at time u. Then the estimator $\hat{X}_i(u)$, conditional on $\{\alpha_i, t_i(u), Y_i(u) = 1, Z_i\}$, is normally distributed with mean $X_i(u) = \alpha_{0i} + \alpha_{1i}u$ and variance $\sigma^2\theta_i(u)$, the usual variance of the estimated mean $\hat{X}_i(u)$ at u using data up to and including u, which depends on timing of measurements for i up to and including u. For $X_i(u) = \alpha_{0i} + \alpha_{1i}u$, $\theta_i(u) = 1/m_{i,u} + (u - \bar{t}_{i,u})^2/SS_{i,u}$, where $t_i(u)$ contains $m_{i,u}$ time-points t_{ij} with mean $\bar{t}_{i,u}$, $SS_{i,u} = \sum_{j=1}^{m_{i,u}}(t_{ij} - \bar{t}_{i,u})^2$.

At any time u, given that i is at risk at time u so that $Y_i(u) = 1$, random effects α_i, longitudinal measurements taken up to and including time u at times $t_i(u)$, and time-independent covariates Z_i, the conditional density for $\{dN_i(u) = r, \hat{X}_i(u) = x\}$ is

$$\text{pr}\{dN_i(u) = r | Y_i(u) = 1, \hat{X}_i(u) = x, \alpha_i, Z_i, t_i(u)\} \times \text{pr}\{\hat{X}_i(u) = x | Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\},$$

which equals

$$\frac{[\lambda_0(u)du\exp\{\gamma X_i(u)+\eta^T Z_i\}]^r[1-\lambda_0(u)du\exp\{\gamma X_i(u)+\eta^T Z_i\}]^{1-r}}{\{2\pi\sigma^2\theta_i(u)\}^{\frac{1}{2}}}\exp\left[-\frac{\{x-X_i(u)\}^2}{2\sigma^2\theta_i(u)}\right];$$

thus, the conditional likelihood of $\{dN_i(u), \hat{X}_i(u)\}$ given $\{Y_i(u) = 1, \alpha_i, Z_i, t_i(u)\}$, up to order du, is

$$[\lambda_0(u)du\exp\{\gamma X_i(u) + \eta^T Z_i\}]^{dN_i(u)}\frac{\exp[-\{\hat{X}_i(u) - X_i(u)\}^2/\{2\sigma^2\theta_i(u)\}]}{\{2\pi\sigma^2\theta_i(u)\}^{\frac{1}{2}}}$$

$$= \exp\left[X_i(u)\left\{\gamma dN_i(u) + \frac{\hat{X}_i(u)}{\sigma^2\theta_i(u)}\right\}\right]\frac{\{\lambda_0(u)\exp(\eta^T Z_i)du\}^{dN_i(u)}}{\{2\pi\sigma^2\theta_i(u)\}^{\frac{1}{2}}}\exp\left\{-\frac{\hat{X}_i^2(u) + X_i^2(u)}{2\sigma^2\theta_i(u)}\right\}.$$

This representation implies that, conditional on $Y_i(u) = 1$,

$$S_i(u, \gamma, \sigma^2) = \gamma\sigma^2\theta_i(u)dN_i(u) + \hat{X}_i(u)$$

is a complete sufficient statistic for α_i, suggesting that, at each time u, conditioning on $S_i(u, \gamma, \sigma^2)$ would remove the dependence of the conditional distribution on the random effects α_i. Then, the conditional intensity process defined as

$$\lim_{du\to 0} du^{-1}\mathrm{pr}\{dN_i(u) = 1|S_i(u, \gamma, \sigma^2), Z_i, t_i(u), Y_i(u)\}$$

is equal to $\lambda_0(u)\exp\{\gamma S_i(u, \gamma, \sigma^2) - \gamma^2\sigma^2\theta_i(u)/2 + \eta^T Z_i\}Y_i(u)$. Reasoning underlying the conditional score estimator follows by analogy with that for estimators for the proportional hazards model with no measurement error.

The conditional intensity of $dN(u) = \Sigma_{j=1}^n dN_j(u)$, given $\{S_i(u, \gamma, \sigma^2), Z_i, t_i(u), Y_i(u), i = 1, ..., n\}$, is $\lambda_0(u)E_0(u, \gamma, \eta, \sigma^2)$, where $E_0(u, \gamma, \eta, \sigma^2) = \sum_{j=1}^n E_{0j}(u, \gamma, \eta, \sigma^2)$,

$$E_{0j}(u, \gamma, \eta, \sigma^2) = \exp\{\gamma S_j(u, \gamma, \sigma^2) - \gamma^2\sigma^2\theta_j(u)/2 + \eta^T Z_j\}Y_j(u).$$

This suggests that a reasonable estimator for $d\Lambda_0(u) = \lambda_0(u)du$ is given by

$$d\hat{\Lambda}_0(u) = dN(u)/E_0(u, \gamma, \eta, \sigma^2).$$

By analogy with the usual score equations derived from the partial likelihood in a proportional hazard model, (γ, η) can be obtained by solving the $(q+1) \times 1$ set of estimating equations

$$\sum_{i=1}^{n} \int \{S_i(u, \gamma, \sigma^2), Z_i^T\}^T \{dN_i(u) - E_{0i}(u, \gamma, \eta, \sigma^2) d\hat{\Lambda}_0(u)\} = 0,$$

which upon substitution of $d\hat{\Lambda}_0(u)$ for $d\Lambda_0(u)$, may be written as

$$\sum_{i=1}^{n} \int \{S_i(u, \gamma, \sigma^2), Z_i^T\}^T \left\{ dN_i(u) - \frac{dN(u) E_{0i}(u, \gamma, \eta, \sigma^2)}{E_0(u, \gamma, \eta, \sigma^2)} \right\} = 0 \tag{2.12}$$

Defining $E_{1j}(u, \gamma, \eta, \sigma^2) = \{S_j(u, \gamma, \sigma^2), Z_j^T\}^T \exp\{\gamma S_j(u, \gamma, \sigma^2) - \gamma^2 \sigma^2 \theta_j(u)/2 + \eta^T Z_j\} Y_j(u)$,

$$E_1(u, \gamma, \eta, \sigma^2) = \sum_{j=1}^{n} E_{1j}(u, \gamma, \eta, \sigma^2),$$

and interchanging the sums in (2.12), the estimating equations are expressed as

$$\sum_{i=1}^{n} \int \left[\{S_i(u, \gamma, \sigma^2), Z_i^T\}^T - \frac{E_1(u, \gamma, \eta, \sigma^2)}{E_0(u, \gamma, \eta, \sigma^2)} \right] dN_i(u) = 0. \tag{2.13}$$

With no measurement error, $\sigma^2 = 0$, (2.13) is identical to the score equations for the maximum partial likelihood estimator of Cox (1975). With $X_i(u)$ time-independent and σ^2 known, the equations are asymptotically equivalent to those proposed by Nakamura (1992). There is an alternative semiparametric estimator with time-independent covariates studied by Buzas (1998).

There are more recent literatures on the selection model. Ribaudo, Thompson &

Allen-Mersh (2000) proposed the application of a random effect selection model in the form of a trivariate Normal model for the joint analysis of QoL response and log survival time. The trivariate Normal model presented by Schluchter is a model that has been discussed in the context of drop-out. This model is a random effect selection model that assumes that the random parameters of a subject's underlying response profile such as intercept and slope of QoL response over time, and the logarithm of the survival time follow a trivariate Normal distribution.

Xu & Zeger (2001) developed latent variable models for joint analysis of longitudinal data comprising repeated measures and times to events, starting with the latent variable formulation of Fawcett and Thomas(1996), and extending and adapting it to the problem of identifying whether a longitudinal variable Y is a useful auxiliary or surrogate variable for event time T given other covariates. The linking linear predictor of Y and T was assumed to follow a Gaussian stochastic process suggested by Diggle (1988).

Song, Davidian, & Tsiatis (2002) assumed that the random effects have distribution in a plausible class with smooth densities, in mixed effects model for longitudinal covariates process belonging to proportional hazards model of event time. They used a class of smooth densities studied by Gallant & Nychka (1987). One speculation of Song *et al.*(2002) is that it is possible that the likelihood based approach using normality yields consistent estimator even when normality is a mis-specification under certain 'nice' conditions through their simulations.

Zeng & Cai (2005) provided the rigorous proof of the consistency of the maximum likelihood estimators and derivation of their asymptotic distributions because there was a lack of theoretical justification of the asymptotic properties for the MLEs even if the ML estimation has been extensively used in the joint analysis of repeated measurements and survival time and has been shown to perform well in numerical studies (Hu, Tsi-

atis & Davidian 1998). Their theoretical results further confirmed that nonparametric maximum likelihood estimation, which was proposed in the literature (Wu & Carroll, 1988; Tsiatis, DeGruttola and Wulfsohn, 1995; Wulfsohn & Tsiatis, 1997), provided efficient estimation. Additionally, it was also shown that the profile likelihood function can be used to give a consistent estimator for the asymptotic variance of the regression coefficients.

Tseng, Hsieh, & Wang (2005) proposed the joint modeling of longitudinal covariates and survival time using accelerated failure time since the accelerated failure time model is an attractive alternative to the Cox model when the proportionality assumption is not appropriate to describe the relationship between the survival time and longitudinal covariates. Hsieh, Tseng, & Wang (2006) recently studied maximum likelihood approach for the joint modelling of survival time and longitudinal covariates in details more.

Song & Wang (2007) proposed semiparametric approaches for joint modeling of longitudinal covariates and survival data with time-varying coefficients. To deal with covariate measurement error, they proposed a local corrected score estimator and a local conditional score estimator which are semiparametric methods in the sense that there is no distributional assumption needed for the underlying true covariates. Li, Wang, & Wang (2007) proposed score functions, named generalized sufficient and conditional scores, for the joint models of a primary endpoint and multiple longitudinal covariate processes by adjusting the bias resulted from the approaches by Li, Zhang & Davidian (2004).

2.3.2 Simultaneous model of failure time and longitudinal data

In many biomedical studies, it is often of interest to investigate the simultaneous effect of treatment or other factors on both patients' quality of life and risk of death. Xu & Zeger

(2001b) and Zeng & Cai (2005) proposed similar simultaneous models of continuous longitudinal outcome $\boldsymbol{Y}$ and survival time T. However, while in the model by Xu & Zeger a common latent process is shared by both $\boldsymbol{Y}$ and T, Zeng & Cai allow individual random effects to affect quality of life and survival time very differently.

In the approach by Zeng & Cai (2005), quality of life and survival time are modeled through parametric and semiparametric models, respectively, assuming a linear mixed effect model for the longitudinal outcomes of quality of life and a multiplicative hazard model for survival time. In both models, observed covariates, which are included as predictors, are assumed to be either time-independent or external time-dependent variables. Unobserved factors enter the models as subject-specific random effects so as to account for unobserved heterogeneity.

For subject i given $T > t$ and the observed history till time t, the longitudinal outcome of quality of life $Y_i(t)$ at time t follows the linear mixed effect model,

$$Y_i(t) = \boldsymbol{X}_i(t)\boldsymbol{\beta} + \tilde{\boldsymbol{X}}_i(t)\boldsymbol{a}_i + \epsilon_i(t),$$

where $\boldsymbol{X}_i(t)$ and $\tilde{\boldsymbol{X}}_i(t)$ are the row vectors of the observed covariates and can be completely different or share some components, $\epsilon_i(t)$ is a white noise process with mean zero and variance σ_y^2, and $\boldsymbol{a}_i$ denotes a vector of subject-specific random effect of dimension k_0 following a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_a$, and $\boldsymbol{\beta}$ is a column vector of coefficients for $\boldsymbol{X}_i(t)$. The random effect $\boldsymbol{a}_i$ reflects the unobserved heterogeneity and is allowed to differ for different levels of covariates $\tilde{\boldsymbol{X}}_i(t)$.

For the survival time T_i given the observed covariates, the observed history till time t, and random effect $\boldsymbol{a}_i$, the conditional hazard rate function is assumed to follow a

multiplicative hazards model,

$$\lambda(t)\exp\{\tilde{\boldsymbol{W}}_i(t)(\boldsymbol{\phi}\circ\boldsymbol{a}_i)+\boldsymbol{W}_i(t)\boldsymbol{\gamma}\},$$

where $\boldsymbol{W}_i(t)$ and $\tilde{\boldsymbol{W}}_i(t)$ are the row vectors of the observed covariates and may share the same components, $\boldsymbol{\phi}$ is a vector of parameters, and $\lambda(t)$ is the baseline hazard rate function, and $\boldsymbol{\gamma}$ is a column vector of coefficients for $\boldsymbol{W}_i(t)$. For the dependence parameter $\boldsymbol{\phi}$ between quality of life and survival time, $\boldsymbol{\phi}=0$ means the dependence can be fully attributed to the observed covariates, and $\boldsymbol{\phi}\neq 0$ implies that such dependence may also be due to some latent variables.

Supposing the survival time is possibly right censored with completely random right-censored time C_i, and assuming N_i, the number of the observed quality of life measurements for subject i, to be non-informative about parameters of interest, the observed data from n subjects are

$$(N_i, Y_i^j, \boldsymbol{X}_i^j, \tilde{\boldsymbol{X}}_i^j), \qquad j=1,\ldots,N_i, i=1,\ldots,n,$$

$$(Z_i, \Delta_i, \{(\boldsymbol{W}_i(t), \tilde{\boldsymbol{W}}_i(t)) : t\le Z_i\}), \qquad i=1,\ldots,n,$$

where for subject i, $(Y_i^j, \boldsymbol{X}_i^j, \tilde{\boldsymbol{X}}_i^j)$ is the j-th observation of $(\boldsymbol{Y}_i, \boldsymbol{X}_i, \tilde{\boldsymbol{X}}_i)$, $Z_i = min(T_i, C_i)$, and $\Delta_i = I(T_i \le C_i)$. Interests are estimating and making inference on the parameters $\boldsymbol{\theta}=(\sigma_y, \boldsymbol{\Sigma}_a, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$ and the baseline cumulative hazard function $\Lambda(t)=\int_0^t \lambda(s)ds$.

Their estimation approach is likelihood-based. In the maximum likelihood method, given the random effects for the i-th subject, the observed quality of life with a multivariate Gaussian distribution is independent of the observed survival time with proportional hazard assumption, and the likelihood contribution of the i-th subject is integrated over the random effects in the joint models. Then, the observed likelihood

function for $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ is expressed as

$$
\begin{aligned}
L \;=\; \prod_{i=1}^{n} \int_{\boldsymbol{a}} \Bigg[& (2\pi\sigma_y^2)^{-N_i/2} \exp\{-(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \tilde{\boldsymbol{X}}_i\boldsymbol{a})^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \tilde{\boldsymbol{X}}_i\boldsymbol{a})/2\sigma_y^2\} \\
& \lambda(Z_i)^{\Delta_i} \exp\left\{ \Delta_i(\tilde{\boldsymbol{W}}_i(Z_i)(\boldsymbol{\phi}\circ\boldsymbol{a}) + \boldsymbol{W}_i(Z_i)\gamma) - \int_0^{Z_i} e^{\tilde{\boldsymbol{W}}_{i(s)}(\boldsymbol{\phi}\circ\boldsymbol{a})+\boldsymbol{W}_{i(s)}\gamma} d\Lambda(s) \right\} \\
& (2\pi)^{-k_0/2}|\boldsymbol{\Sigma}_a|^{-1/2} \exp\{-\boldsymbol{a}^T\boldsymbol{\Sigma}_a^{-1}\boldsymbol{a}/2\} \Bigg] d\boldsymbol{a},
\end{aligned}
$$

where $\boldsymbol{Y}_i$ denotes the vector of $(Y_i^1, ..., Y_i^{N_i})^T$, $\boldsymbol{X}_i$ denotes the matrix of $((\boldsymbol{X}_i^1)^T, ..., (\boldsymbol{X}_i^{N_i})^T)^T$, $\tilde{\boldsymbol{X}}_i$ denotes $((\tilde{\boldsymbol{X}}_i^1)^T, ..., (\tilde{\boldsymbol{X}}_i^{N_i})^T)^T$, and k_0 is the dimension of $\boldsymbol{a}$.

EM algorithms are employed for the maximum likelihood estimates for $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ over a set in which $\boldsymbol{\theta}$ is in a bounded set and Λ belongs to a space consisting of all the increasing functions with $\Lambda(0) = 0$. It is clear that the maximum likelihood estimate for Λ can be chosen as a step function with jumps only at the observed failure times. In the EM algorithm, $\boldsymbol{a}_i$ is considered as the missing statistics for $i = 1, \ldots, n$. Therefore, the M-step solves the conditional score equation from the complete data given the observations, where the conditional expectation can be evaluated in the E-step. The iteration between E-step and M-step is conducted until the estimates converge. The final maximum likelihood estimate for $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ is denoted by $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Lambda}})$.

The variance estimator for $\hat{\boldsymbol{\theta}}$ is obtained by using the profile likelihood function whose logarithm is defined as $pl_n(\boldsymbol{\theta}) = \max_\Lambda n^{-1} \sum_{i=1}^n q_i(\boldsymbol{\theta}, \Lambda)$ where $q_i(\boldsymbol{\theta}, \Lambda)$, $i = 1, \ldots, n$, is the logarithm of the observed likelihood function for the i-th subject. Particularly, an efficient algorithm, which is based on the EM-algorithm to calculate the profile likelihood function, is proposed and called as the PEME algorithm (partial expectation, maximization and evaluation).

2.4 Penalized Quasi-Likelihood Approach

In the view of the cumbersome and often intractable numerical integrations required for a full likelihood analysis, several suggestions were made for approximate inference in generalized linear mixed models and other nonlinear variance component models. One approach was proposed by Breslow & Clayton (1993) with some modifications to a Laplace expansion in order to motivate standard estimating equations that may be solved by iterative application of normal theory variance components procedures. In this section, we mainly review the penalized quasi-likelihood for generalized linear mixed model proposed by Breslow & Clayton (1993), and the bias correction in the penalized quasi-likelihood estimators proposed by Breslow & Lin (1995).

2.4.1 Penalized quasi-likelihood in generalized linear mixed model

The penalized quasi-likelihood (PQL) method exploited by Green (1987) for semiparametric regression analysis is available for inference in hierarchical models where the focus is on shrinkage estimation of the random effects (Robinson 1991). The PQL was proposed as an approximate Bayes procedure for some commonly occurring GLMM's by Laird (1978). Breslow & Clayton (1993) considered two closely related approximate methods (Penalized Quasi-Likelihood and Marginal Quasi-Likelihood) of inference in GLMM's and investigated their suitability for practical work by means of Monte Carlo studies and illustrative applications. Here we review only the PQL in their paper. They provided the PQL criterion motivated by approximating the integrated quasi-likelihood, and developed an approximate GLM for the marginal distribution of the data. The approximate GLM is related to the generalized estimating equation approach of Zeger et al. (1988). They note that PQL tends to underestimate somewhat the variance com-

ponents and (in absolute value) fixed effects when applied to clustered binary data, but the situation improves rapidly for binomial observations having denominators greater than one.

Within the framework of the generalized linear mixed model(GLMM), given an unobserved vector of random effects, observations are assumed to be conditionally independent with means that depend on the linear predictor through a specified link function and conditional variances that are specified by a variance function, known prior weights and a scale factor. The random effects are assumed to be normally distributed with mean zero and dispersion matrix depending on unknown variance components.

Consider hierarchical model and denote y_i, $i = 1, \ldots, n$, as the i-th observation of a univariate response variable with two vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ of explanatory variables associated with the fixed and random effects respectively. The n responses may be blocked in some way, for example when they involve repeated measures on the same subject. Suppose that, given a q-dimensional vector $\boldsymbol{b}$ of random effects, the y_i are conditionally independent with means $\mathrm{E}(y_i|\boldsymbol{b}) = \mu_i^b$ and variances $\mathrm{Var}(y_i|\boldsymbol{b}) = \phi a_i v(\mu_i^b)$, where $v(\cdot)$ is a specified variance function, a_i is a known constant (e.g., the reciprocal of a binomial denominator) and ϕ is a dispersion parameter that may or may not be known. The conditional mean is related to the linear predictor $\eta_i^b = \boldsymbol{x}_i^T\boldsymbol{\alpha} + \boldsymbol{z}_i^T\boldsymbol{b}$ by the link function $g(\mu_i^b) = \eta_i^b$, with inverse $h = g^{-1}$, where $\boldsymbol{\alpha}$ is a p vector of fixed effects. Denoting the observation vector by $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and the design matrices with rows $\boldsymbol{x}_i^T$ and $\boldsymbol{z}_i^T$ by $\boldsymbol{X}$ and $\boldsymbol{Z}$, the conditional mean satisfies

$$\mathrm{E}(\boldsymbol{y}|\boldsymbol{b}) = h(\boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{b}).$$

Assume that $\boldsymbol{b}$ has a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{\theta})$ depending on an unknown vector $\boldsymbol{\theta}$ of variance components. In binomial, Poisson, and hypergeometric specifications for the conditional distribution

of y_i, the dispersion parameter ϕ is fixed at unity. In other cases, however, it may be estimated together with $\boldsymbol{\theta}$ as a parameter in the covariance matrix of the marginal distribution of $\boldsymbol{y}$.

The integrated quasi-likelihood function used to estimate $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ is defined by

$$e^{ql(\boldsymbol{\alpha},\boldsymbol{\theta})} \propto |\boldsymbol{D}|^{-1/2} \int \exp\left[-\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i; \mu_i^b) - \frac{1}{2}\boldsymbol{b}^T \boldsymbol{D}^{-1}\boldsymbol{b} \right] d\boldsymbol{b}, \tag{2.14}$$

where $d_i(y, \mu) = -2\int_y^{\mu} \frac{y-\mu}{a_i v(u)} du$ denotes the deviance measure of fit. If, conditionally on $\boldsymbol{b}$, the observations are drawn from a linear exponential family with variance function $v(\cdot)$, then the deviance is well known to equal to the scaled difference $2\phi\{l(y; y, \phi) - l(y; \mu, \phi)\}$, where $l(y; \mu, \phi)$ denotes the conditional likelihood of y given its mean μ (McCullagh & Nelder 1989). In this case $ql(\boldsymbol{\alpha}, \boldsymbol{\theta})$ represents the true log-likelihood of the data. The primary difficulty in implementing full likelihood inference lies in the integrations needed to evaluate ql and its partial derivatives.

The equation (2.14) can be written as $c|\boldsymbol{D}|^{-1/2} \int e^{-\boldsymbol{\kappa}(\boldsymbol{b})} d\boldsymbol{b}$, and then applied with Laplace's method for integral approximation (Barndorff-Nielsen & Cox 1989; Tierney & Kadane 1986). Let $\boldsymbol{\kappa}'$ and $\boldsymbol{\kappa}''$ denote the q vector and $q \times q$ dimensional matrix of first- and second-order partial derivatives of $\boldsymbol{\kappa}$ with respect to $\boldsymbol{b}$. Ignoring the multiplicative constant c, the approximation yields

$$ql(\boldsymbol{\alpha}, \boldsymbol{\theta}) \approx -\frac{1}{2}\log|\boldsymbol{D}| - \frac{1}{2}\log|\boldsymbol{\kappa}''(\tilde{\boldsymbol{b}})| - \boldsymbol{\kappa}(\tilde{\boldsymbol{b}}), \tag{2.15}$$

where $\tilde{\boldsymbol{b}} = \tilde{\boldsymbol{b}}(\boldsymbol{\alpha}, \boldsymbol{\theta})$ denotes the solution to

$$\boldsymbol{\kappa}' = -\sum_{i=1}^{n} \frac{(y_i - \mu_i^b)\boldsymbol{z}_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} + \boldsymbol{D}^{-1}\boldsymbol{b} = 0$$

that minimizes $\boldsymbol{\kappa}(\boldsymbol{b})$. Differentiating again with respect to $\boldsymbol{b}$, we have

$$\begin{aligned}\boldsymbol{\kappa}'' &= -\sum_{i=1}^{n} \frac{\boldsymbol{z}_i \boldsymbol{z}_i^T}{\phi a_i v(\mu_i^b)[g'(\mu_i^b)]^2} + \boldsymbol{D}^{-1} + \boldsymbol{R} \\ &\approx \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{D}^{-1}, \end{aligned} \tag{2.16}$$

where $\boldsymbol{W}$ is the $n \times n$ diagonal matrix with diagonal terms $w_i = \{\phi a_i v(\mu_i^b)[g'(\mu_i^b)]^2\}^{-1}$ that are recognizable as the GLM iterated weights (Firth 1991, McCullagh & Nelder 1989). The remainder term $\boldsymbol{R} = -\sum_{i=1}^{n}(y_i - \mu_i^b)\boldsymbol{z}_i \frac{\partial}{\partial \boldsymbol{b}}\left[\frac{1}{\phi a_i v(\mu_i^b) g'(\mu_i^b)}\right]$ has expectation 0 and is thus, in probability as a function of n, of lower order than the two leading terms in the equation of $\boldsymbol{\kappa}''$. $\boldsymbol{R}$ equals $\boldsymbol{0}$ for the canonical link functions, for which $g'(\mu) = v^{-1}(\mu)$ (McCullagh & Nelder 1989). Combining (2.14)–(2.16) and ignoring $\boldsymbol{R}$ leads to

$$ql(\boldsymbol{\alpha}, \boldsymbol{\theta}) \approx -\frac{1}{2}\log|\boldsymbol{I} + \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} \boldsymbol{D}| - \frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i, \mu_i^{\tilde{b}}) - \frac{1}{2}\tilde{\boldsymbol{b}}^T \boldsymbol{D}^{-1}\tilde{\boldsymbol{b}}, \tag{2.17}$$

where $\tilde{\boldsymbol{b}}$ is chosen to maximize the sum of the last two terms.

Assuming that the GLM iterative weights vary slowly (or not at all) as a function of the mean, the first term in this expression is ignored, and $\boldsymbol{\alpha}$ is chosen to maximize the second. Thus $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}}) = (\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \hat{\boldsymbol{b}}(\boldsymbol{\theta}))$, where $\hat{\boldsymbol{b}}(\boldsymbol{\theta}) = \tilde{\boldsymbol{b}}(\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}))$, jointly maximize Green's (1987) PQL

$$-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i, \mu_i^b) - \frac{1}{2}\boldsymbol{b}^T \boldsymbol{D}^{-1}\boldsymbol{b}. \tag{2.18}$$

Differentiation with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{b}$ leads to the score equations for the mean parameters:

$$\begin{aligned}\sum_{i=1}^{n} \frac{(y_i - \mu_i^b)\boldsymbol{x}_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} &= 0 \\ \sum_{i=1}^{n} \frac{(y_i - \mu_i^b)\boldsymbol{z}_i}{\phi a_i v(\mu_i^b) g'(\mu_i^b)} &= \boldsymbol{D}^{-1}\boldsymbol{b}. \end{aligned}$$

For the estimation of variance component $\boldsymbol{\theta}$, we substitute the maximized value of (2.18) into (2.17) and evaluate $\boldsymbol{W}$ at $(\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \hat{\boldsymbol{b}}(\boldsymbol{\theta}))$, which generate an approximate profile quasi-likelihood function $ql(\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ for inference on $\boldsymbol{\theta}$. To make degrees-of-freedom adjustments that account for the fact that $\hat{\boldsymbol{\alpha}}$ rather than $\boldsymbol{\alpha}$ appears in the approximate profile quasi-likelihood function $ql(\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \boldsymbol{\theta})$, we modify $ql(\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ to the REML version (Patterson & Thompson 1971) in practice. By differentiating the modified profile quasi-likelihood with respect to the components of $\boldsymbol{\theta}$, we obtain the estimating equations for the variance parameters.

2.4.2 Bias correction in penalized quasi-likelihood

The approach proposed by Breslow & Clayton (1993) have been applied to a wide variety of generalized linear mixed models. Although the approximate procedure have been demonstrated to work reasonably well for discrete data problems with moderate to large cell frequencies, their performance is less satisfactory when the data are sparse. Breslow & Lin (1995) derived the general expressions for the asymptotic biases in approximate estimators of regression coefficients and variance component, for small values of the variance component, in generalized linear mixed models with canonical link function and a single source of extraneous variation. Their numerical studies of a series of matched pairs of binary outcomes showed that the first order estimators of the variance component are seriously biased, and they provided the easily computed correction factors which produce satisfactory estimators of small variance components. Their variance correction factors for a series of matched pairs of binomial observations rapidly approach one as the binomial denominators increase.

Let the data be in a series of m clusters of observations (y_{ij}, x_{ij}), where i identifies the cluster, $j = 1, \ldots, n_i$ identifies subjects within clusters and x_{ij} are p-vectors of explanatory variables associated with the univariate outcome y_{ij}. Given an unobserved

random effect b_i, the observations in the i-th cluster are assumed to have log conditional density

$$l_i(\alpha; b_i) = \sum_{j=1}^{n_i} \frac{a_{ij}}{\phi}\{y_{ij}\eta_{ij} - h(\eta_{ij})\} + c(y_{ij}; \phi), \tag{2.19}$$

where $\eta_{ij} = x_{ij}^T\alpha + b_i$ denotes a linear predictor, the a_{ij} are prior weights, and ϕ is a known scale parameter. This restriction to canonical link functions (McCullagh & Nelder, 1989) implies that the the conditional means $\mu_{ij}^{b_i} = \mathrm{E}(y_{ij}|b_i) = g^{-1}(\eta_{ij})$ and variances $\mathrm{Var}(y_{ij}|b_i) = \phi a_{ij}^{-1} v(\mu_{ij}^{b_i})$ are related via $g' = 1/h'' = 1/v$ for link and variance functions g and v, respectively. The b_i are assumed to be a random sample from a normal population with mean 0 and variance θ. Thus the likelihood for the observed data is

$$L(\alpha, \theta) = \prod_{i=1}^{m} L_i(\alpha, \theta) = \prod_{i=1}^{m} (2\pi\theta)^{-\frac{1}{2}} \int e^{l_i(\alpha,b) - b^2/2\theta} db. \tag{2.20}$$

Denote $(\hat{\alpha}, \hat{\theta})$ as the true maximum likelihood estimator. For approximations, we consider the derivatives $l_i^{(k)} = \partial^k l_i / \partial b^k$. Using Laplace method (e.g. Barndorff-Nielson & Cox 1989), the likelihood function in (2.20) may be approximated by expanding the integrand in a Taylor series about its maximizing value $\tilde{b}_i$, where $\tilde{b}_i = \tilde{b}_i(\alpha, \theta)$ solves $\tilde{b}_i = \theta l_i^{(1)}(\alpha, \tilde{b}_i)$. Setting $\tilde{l}_i^{(k)} = l_i^{(k)}(\alpha, \tilde{b}_i)$, a quartic expansion gives

$$\begin{aligned}
L_i(\alpha, \theta) &\doteq (2\pi\theta)^{-\frac{1}{2}} \exp\left(\tilde{l}_i - \frac{\tilde{b}_i^2}{2\theta}\right) \int \exp\left\{\frac{1}{2}\left(\tilde{l}_i^{(2)} - \frac{1}{\theta}\right)(b - \tilde{b}_i)^2\right\} \\
&\qquad \times \left\{1 + \frac{1}{6}\tilde{l}_i^{(3)}(b - \tilde{b}_i)^3 + \frac{1}{24}\tilde{l}_i^{(4)}(b - \tilde{b}_i)^4\right\} db \\
&= (1 - \theta\tilde{l}_i^{(2)})^{-\frac{1}{2}} \exp\left(\tilde{l}_i - \frac{\tilde{b}_i^2}{2\theta}\right)\left\{1 + \frac{\theta^2\tilde{l}_i^{(4)}}{8(1 - \theta\tilde{l}_i^{(2)})^2}\right\} \\
&\doteq (1 - \theta\tilde{l}_i^{(2)})^{-\frac{1}{2}} \exp\left\{\tilde{l}_i - \frac{\tilde{b}_i^2}{2\theta} + \frac{\theta^2\tilde{l}_i^{(4)}}{8(1 - \theta\tilde{l}_i^{(2)})^2}\right\},
\end{aligned}$$

where we evaluated the integral by taking expectations with respect to a normal variate having mean $\tilde{b}_i$ and variance $\theta/(1 - \theta\tilde{l}_i^{(2)})$. We define the first order Laplace approxi-

mation to the log likelihood using only the leading terms of this expansion,

$$l_{L1}(\alpha,\theta) = \sum_{i=1}^{m}\left\{ -\frac{1}{2}\log(1-\theta\tilde{l}_i^{(2)}) + \tilde{l}_i - \frac{\tilde{b}_i^2}{2\theta}\right\}. \tag{2.21}$$

The Laplace approximation estimator $(\hat{\alpha}_{L1}, \hat{\theta}_{L1})$ are defined to be those that maximize l_{L1}.

Breslow & Clayton (1993), following Green (1987), termed the log conditional likelihoods (2.19) minus the penalty term $\sum_i b_i^2/(2\theta)$ a log penalized quasi-likelihood in recognition of the fact that l_i requires specification only of the mean-variance relationship for the conditional distribution. Maximizing the penalized quasi-likelihood as a function of $b = (b_1, \ldots, b_m)^T$ for fixed (α, θ) leads to an objective function

$$l_p(\alpha,\theta) = \sum_{i=1}^{m}\left(\tilde{l}_i - \frac{\tilde{b}_i^2}{2\theta}\right)$$

that equals the sum of the last two terms in the first order Laplace approximation (2.21). The penalized quasi-likelihood estimator of the regression coefficients is defined to be the value $\hat{\alpha}_P(\theta)$ that maximizes $l_i(\alpha,\theta)$ for fixed θ. The optimization may be programmed as a problem in iterated weighted least squares. Specifically, let $\boldsymbol{Y}$ denote the $N = \sum_i n_i$ dimensional 'working vector' whose components in lexicographic order are $Y_{ij} = x_{ij}^T\alpha + \tilde{b}_i + (y_{ij} - \mu_{ij}^{\tilde{b}_i})/v_{ij}^{\tilde{b}_i}$; let $\boldsymbol{V}$ denote the $N \times N$ block diagonal covariance matrix whose $n_i \times n_i$ dimensional diagonal submatrices V_i have terms $\phi(a_{ij}v_{ij}^{\tilde{b}_i})^{-1} + \theta$ along their diagonals and off-diagonal elements θ; and let $\boldsymbol{X}$ denote the $N \times p$ design matrix with rows x_{ij}^T. Then, the Fisher scoring algorithm for solving the penalized quasi-likelihood equations

$$\frac{\partial l_p(\alpha,\theta)}{\partial\alpha} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{a_{ij}}{\phi}(y_{ij} - \mu_{ij}^{\tilde{b}_i})x_{ij} = 0 \tag{2.22}$$

for α reduces to iterative solution of $(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})\alpha = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$ (Green, 1987).

Estimation of θ under penalized quasi-likelihood treats the working vector $\boldsymbol{Y}$ as normally distributed with covariance matrix $\boldsymbol{V}$ depending on θ, except that the dependence of the terms $v_{ij}^{\tilde{b}_i}$ on θ through $\tilde{b}_i$ is ignored when calculating derivatives (Breslow & Clayton, 1993). The principal advantage of this approach over the Laplace approximations is that it may be implemented using standard software for mixed model analysis. In this paper by Breslow & Lin (1995) the simpler maximum likelihood is used since they focused on asymptotic results rather than small sample properties while Breslow & Clayton (1993) used the restricted maximum likelihood normal theory approach. Thus, the penalized quasi-likelihood variance estimating equation is

$$\begin{aligned}\tilde{U}(\theta) &= \frac{1}{2}\left\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\alpha})^T\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial\theta}\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\alpha}) - tr\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial\theta}\right)\right\}\Bigg|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_P(\theta)} \\ &= \frac{1}{2}\sum_{i=1}^{m}\left(\tilde{l}_i^{(1)2} + \frac{\tilde{l}_i^{(2)}}{1-\theta\tilde{l}_i^{(2)}}\right)\Bigg|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_P(\theta)} = 0. \qquad (2.23)\end{aligned}$$

The penalized quasi-likelihood estimators $(\hat{\boldsymbol{\alpha}}_P, \hat{\theta}_P)$ simultaneously solve equations (2.22) and (2.23). While $\hat{\boldsymbol{\alpha}}_P(\theta)$ maximizes $l_P(\alpha, \theta)$, however, $\hat{\theta}_P$ does not maxmize $l_P\{\hat{\boldsymbol{\alpha}}_P(\theta), \theta\}$.

Depending upon the distribution of the data and thus the link function in canonical generalized linear mixed models, the estimates of regression coefficients may be heavily influenced by the value assumed for the dispersion parameter. Accordingly, since some of the bias in an estimator of α may arise from bias in the corresponding estimator of θ, Breslow & Lin (1995) studied the bias in the estimator of α for small fixed θ, and then the bias in the estimator of θ.

First, we expand the true log-likelihood and the approximation in Taylor series

about $\theta = 0$. Then, we have

$$\begin{aligned} l &= \sum_{i=1}^{m} l_{i0} + \theta \sum_{i=1}^{m} \left(\frac{l_{i0}^{(2)}}{2} + \frac{l_{i0}^{(1)2}}{2} \right) + \frac{\theta^2}{2} \theta \sum_{i=1}^{m} \left(\frac{l_{i0}^{(2)2}}{2} + l_{i0}^{(1)2} l_{i0}^{(2)} + l_{i0}^{(1)} l_{i0}^{(3)} + \frac{l_{i0}^{(4)}}{4} \right) + o(\theta^2) \\ l_P &= l - \frac{\theta}{2} \sum_{i=1}^{m} l_{i0}^{(2)} - \frac{\theta^2}{4} \sum_{i=1}^{m} l_{i0}^{(2)2} - \frac{\theta^2}{2} \sum_{i=1}^{m} l_{i0}^{(1)} l_{i0}^{(3)} - \frac{\theta^2}{8} \sum_{i=1}^{m} l_{i0}^{(4)} + o(\theta^2), \end{aligned}$$

where we use the fact that

$$\left. \frac{\partial \tilde{b}_i}{\partial \theta} \right|_{\theta=0} = l_{i0}^{(1)}, \qquad \left. \frac{\partial \tilde{b}_i}{\partial \theta^2} \right|_{\theta=0} = 2 l_{i0}^{(1)} l_{i0}^{(2)}.$$

Then, the difference between $\hat{\alpha}_P$ and $\hat{\alpha}$ are studied by expanding

$$0 = \left. \frac{\partial l}{\partial \boldsymbol{\alpha}} \right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} = \left. \frac{\partial l}{\partial \boldsymbol{\alpha}} \right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_P} + \left. \frac{\partial^2 l}{\partial \boldsymbol{\alpha} \boldsymbol{\alpha}^T} \right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*} (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_P).$$

Consequently, we have

$$\hat{\boldsymbol{\alpha}}_P = \hat{\boldsymbol{\alpha}} + \frac{\theta}{2} (\boldsymbol{X}^T \boldsymbol{W}_0 \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{u} + o(\theta), \tag{2.24}$$

where $\boldsymbol{W}_0$ denotes the diagonal matrix with weight $a_{ij} v_{ij}^0 / \phi$ on the diagonal $\boldsymbol{u}$ is an $N \times 1$ vector with components $a_{ij} v(\mu_{ij}^0) v'(\mu_{ij}^0) / \phi$ and both $\partial l / \partial \boldsymbol{\alpha}$ and $\partial l_P / \partial \boldsymbol{\alpha}$ are evaluated at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_P(\theta)$. The corrected penalized quasi-likelihood estimate is obtained by subtracting the linear term in (2.24) from $\hat{\boldsymbol{\alpha}}_P$.

The asymptotic biases in the estimator of θ derived from the penalized quasi-likelihood were evaluated by equating expansions of the log profile likelihood $l^\sharp(\theta) = \log L\{\hat{\boldsymbol{\alpha}}(\theta), \theta\}$ to expansion of the penalized quasi-likelihood approximations. Then, we have

$$\frac{\hat{\theta}_P}{\hat{\theta}} = \left(\frac{\partial^2 l_P^\sharp}{\partial \theta^2} \right)^{-1} \left. \frac{\partial^2 l^\sharp}{\partial \theta^2} \right|_{\theta=0} \sim \frac{B - C}{C},$$

where

$$
\begin{aligned}
B &= \sum_i l_{i0}^{(2)2}/2 - \boldsymbol{u}^T \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{W}_0 \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{u}/4, \\
C &= \sum_i l_{i0}^{(4)}/4, \\
D &= \sum_i l_{i0}^{(2)2}/2.
\end{aligned}
$$

Lin & Brelsow (1996) also derived the biases correction in generalized linear mixed models with multiple components of dispersion.

Chapter 3

JOINT ANALYSIS FOR SURVIVAL TIME AND LONGITUDINAL CATEGORICAL MEASUREMENTS OF QUALITY OF LIFE IN HEAD AND NECK CANCER PATIENTS

3.1 Introduction

When choosing a treatment, it is well known that decision–making on treatment is frequently based on probability of survival. However, when there are multiple treatment modalities with similar survival rates, Quality of Life (QoL) factors are raised as important considerations for patients. In particular, oncology community has recognized that QoL and functional status are the major outcome variables in the evaluation of head and neck cancer treatment because of the potential impact on critical functions such as speech, swallowing, and breathing, as well as cosmesis and communication.

Many studies have been conducted for QoL without considering the survival time. For example, Terrell *et al.* (2004) investigated clinical predictors of QoL in a large intervention study of patients with head and neck cancer. Ringash, Bezjak, O'Sullivan

and Redelmeier (2004) studied QoL of particulary laryngeal cancer patients among those with head and neck cancer. Holloway *et al.* (2005) studied psychosocial effects in long-term head and neck cancer survivors. Fang *et al.* (2004) studied changes in QoL of head and neck cancer patients following postoperative radiotherapy. Most recently, Nibu *et al.* (2010) collected QoL data at scheduled clinic appointments of head and neck cancer patients and conducted a longitudinal QoL analysis. All these studies did not take the survival time into consideration. In order to completely understand the factors influencing both QoL and survival, it is important to study the QoL and survival simultaneously.

The Carolina Head and Neck Cancer Study (CHANCE) is a population based epidemiologic study conducted at 60 hospitals in 46 counties in North Carolina from 2002 through 2006 (Divaris *et al.* 2010). Patients were diagnosed with head and neck cancer (oral, pharynx, and larynx cancer) from 2002–2006. Their survival status was collected up to 2007 and QoL was evaluated over time for three years after diagnosis. QoL information was collected through questionnaires. Based on summary scores of the five domains of self-perceived quality of life including Physical Well-Being (PWB), Social/Family Well-Being (SWB), Emotional Well-Being (EWB), Functional Well-Being (FWB) and Head and Neck Cancer Specific symptoms (HNCS), patient's QoL information was classified into satisfaction or dissatisfaction with life. Survival time is defined as the time to death from diagnosis. Demographic and life style characteristics, medical histories and clinical factors are also collected. It is of interest to elucidate the variables which are associated with both QoL satisfaction and survival time for patients with head and neck cancer. Additionally, the longitudinal QoL satisfaction outcomes and survival time are correlated within a patient, and this dependency should be taken into account in the analysis.

Among the existing approaches for longitudinal data and survival time, the selec-

tion model and the pattern mixture model have been widely used. The selection model estimates the distribution of survival time given longitudinal data. The selection model with continuous longitudinal data was studied by Tsiatis, Degruttola, and Wulfsohn (1995), Wulfsohn and Tsiatis (1997), Henderson, Diggle and Dobson (2000), Tsiatis and Davidian (2001), Song, Davidian and Tsiatis (2002), Tseng, Hsieh and Wang (2005) and Song and Wang (2007). The selection model with categorical longitudinal data was considered by Faucett, Schenker and Elashoff (1998), Huang *et al.* (2001), Xu and Zeger (2001a,b) and Larsen (2004). The pattern mixture model focuses on the trend of longitudinal outcomes conditional on survival time. The pattern mixture model with continuous longitudinal outcomes was studied by Wu and Carroll (1988), Wu and Bailey (1989), Hogan and Laird (1997), Ribaudo, Thompson and Allen-Mersh (2000) and more recently Ding and Wang (2008). Albert and Follmann (2000) proposed to model repeated count data subject to informative dropout and Albert, Follmann, Wang and Suh (2002) and Albert and Follmann (2007) studied binary longitudinal data with informative missingness. These methods cannot be applied directly to assess covariate effects on both outcomes. Simultaneous modeling of the longitudinal and survival data are needed for such purpose. Xu and Zeger (2001b) and Zeng and Cai (2005a) proposed simultaneous models of continuous longitudinal outcome and survival time. In their articles, heterogeneity caused by unobserved factors is represented using subject-specific random effects. Given random effects, survival time and the repeated measurements of longitudinal outcomes are assumed to follow a Cox proportional hazards model and a Gaussian distribution, respectively. Recently, Elashfoff, Li and Ni (2007, 2008) proposed a more general joint model which incorporates a competing risks model for survival endpoint. Rizopoulos, Verbeke and Molenberghs (2008) assumed an accelerated failure time model and proposed to consider two separate sets of random effects for the continuous longitudinal process and survival time process, linking them

using a copula function. As an extension of this study, Rizopoulos, Verbeke, Lesaffre and Vanrenterghem (2008) considered longitudinal binary data with excess zeros and proposed a two-part shared parameter model framework. In the Bayesian perspective, Wang and Taylor (2001) and Brown and Ibrahim (2003) studied the simultaneous analysis of continuous longitudinal outcomes and survival time. Hu, Li and Li (2009) extended the existing Bayesian approach by considering the more general joint model of Elashfoff *et al.* (2008) with multiple types of failures in the failure time data.

Compared to the studies for continuous longitudinal data and survival time, relatively little work has been done in the joint modeling frame work for categorical longitudinal data and survival time. However, the outcomes may not be continuous in some biomedical studies, for example, where the outcomes are disease symptom with categories of mild/moderate/severe, quality of life measurements with dissatisfied/satisfied, or dichotomized test results with categories of positive/negative. With these categorical longitudinal outcomes, the existing theory cannot be applied directly and the numerical algorithm needs to be modified. Therefore, in this paper, we investigate the simultaneous modeling of survival time and longitudinal categorical outcomes. Furthermore, hazards model for survival time is extended to allow multiple strata in our approach. Random effects are introduced into the proposed models to account for the dependence between survival time and longitudinal outcomes due to unobserved factors.

The outline of this paper is as follows. We begin by describing the details of the CHANCE study in Section 3.2. In Section 3.3, we then present a simultaneous modeling for longitudinal categorical outcomes and survival time, and describe the inference procedure. Asymptotic properties of the proposed estimators and the technical details of their proofs are given in Section 3.4 and Section 3.5, respectively. Numerical results from simulation studies are given in Section 3.6. The analysis of the CHANCE study

is provided in Section 3.7. In Section 3.8, we discuss some further consideration and generalization.

3.2 The CHANCE Study

The Carolina Head and Neck Cancer Study (CHANCE) is the largest epidemiologic study of squamous cell carcinoma of the head and neck in the United States and the first to include a significant number of black patients. Patients who were diagnosed with head and neck cancer (oral, pharynx, and larynx cancer) from 2002 to 2006 were evaluated for Quality of Life (QoL) at maximum three times over follow-up at one to six months, one year and three years after diagnosis. At each evaluation, they were given questionnaires asking about their QoL satisfaction. Ending in December 2009, information on QoL has been obtained from 587 head and neck cancer patients. Based on the death information through 2007 available from the National Death Index (NDI), 91 patients died. It is of interest to study the effects of demographic and life style characteristics, medical histories, and clinical factors on patients' QoL and survival time. In particular, it is of interest to compare between African-Americans and Whites since it is known that African-Americans have a higher incidence of head and neck cancer and worse survival than Whites. Furthermore, because QoL outcomes are especially critical for physicians, head and neck cancer patients, and their caregivers, more research was needed on the experiences of survivors, especially among black patients. Given the paucity of data and studies on QoL among African-American head and neck cancer survivors, this study yields valuable new data.

To collect QoL information, the Functional Assessment of Cancer Therapy–Head and Neck Version 4 (FACT–H&N) series of questionnaires was used (Cella *et al.* 1993; Cella 1994; D'Antionio, Zimmerman, Cella and Long 1996; List *et al.* 1996). This QoL instrument was specifically designed for use of head and neck cancer patients and

consists of five primary QoL domains: Physical Well-Being (PWB), Social/Family Well-Being (SWB), Emotional Well-Being (EWB), Functional Well-Being (FWB) and Head and Neck Cancer Specific symptoms (HNCS). FACT-HN is the overall measurement of QoL incorporating all these domains. Each question has 5 scales of QoL measurement. Among them, the 3 high levels of "very much", "quite a bit" and "somewhat" are categorized to "satisfied" and the other 2 low levels of "a little bit" and "not at all" belong to "dissatisfied". We are interested in identifying the factors associated with both QoL and survival time. Longitudinal QoL outcomes are binary measurements with 1 ("satisfied") and 0 ("dissatisfied") on the five QoL domains, and survival time is the time to death from diagnosis. In this study, 33 among 587 patients were excluded due to missing data on household income, beer and QoL information resulting in 554 patients in the analysis. Eighty-five patients deceased by the end of 2007 and the censoring rate is 85%. Table 3.1 shows the descriptive statistics of predictors: demographics factors – race, household income, age at diagnosis, number of persons supported by household income; alcohol factor – the number of 12 oz. beers consumed per week; medical history factors – BMI and total number of medical conditions reported; treatment history factors – surgery, radiation therapy and chemotherapy; primary tumor data factors – primary tumor site and stage; time from diagnosis to each QOL survey. We analyze a QoL domain of the Head and Neck Cancer Specific symptoms (HNCS) and survival time, and Table 3.2 shows the descriptive statistics of outcome variables: longitudinal HNCS binary outcomes at three surveys, survival time from diagnosis and censorship indicator. The number of observations per patient ranges 1 to 3 with average of 1.93. We are interested in investigating factors which are associated with QoL and survival. In the next section, we formulate a general model and propose an inference procedure.

Table 3.1: Descriptive statistics of predictors in the CHANCE study

Categorical variables	Freq.	%
Total	554	100.00
Race		
– White	444	80.14
– African-American	110	19.86
Household income		
– level1: 0–10K	157	28.86
– level2: 20–30K	129	23.71
– level3: 40–50K	107	19.67
– level4: ≥ 60K	151	27.76
# of 12 oz. beers consumed per week		
– None	103	18.59
– less than 1	50	9.03
– 1 to 4	94	16.97
– 5 to 14	129	23.29
– 15 to 29	69	12.45
– 30 or more	109	19.68
Surgery		
– No	237	42.78
– Yes	317	57.22
Radiation therapy		
– No	131	23.65
– Yes	423	76.35
Chemotherapy		
– No	324	58.48
– Yes	230	41.52
Tumor site		
– Oral & Pharyngeal	346	62.45
– Laryngeal	208	37.55
Tumor stage		
– I	144	25.99
– II	93	16.79
– III	99	17.87
– IV	218	39.35

Continuous variables	n	mean	std.dev	min	median	max
Age at diagnosis	554	59.11	10.19	24.00	59.00	80.00
# of persons supported by household income	554	2.23	1.06	1.00	2.00	5.00
BMI	554	27.47	5.98	15.66	26.48	56.28
Total # of medical conditions reported	554	.92	1.10	.00	1.00	6.00
Time at 1st survey measurement (years)	209	.41	.45	.09	.28	3.55
Time at 2nd survey measurement (years)	500	1.85	.86	.44	1.81	3.91
Time at 3rd survey measurement (years)	353	3.49	.54	1.88	3.54	4.88

– Time at survey measurement is from diagnosis.

Table 3.2: Descriptive statistics of outcome variables in the CHANCE study

Longitudinal QoL binary outcomes	1st survey Freq.	1st survey %	2nd survey Freq.	2nd survey %	3rd survey Freq.	3rd survey %
HNCS	209	100.00	500	100.00	352	100.00
– Dissatisfied (=0)	81	38.76	120	24.00	72	20.45
– Satisfied (=1)	128	61.24	380	76.00	280	79.55

Survival outcomes	n	mean	std.dev	min	median	max
min(Survival time, Censored time) (years)	554	3.07	1.04	.44	2.91	5.98
	Freq.	%				
Censorship	554	100.00				
– Alive	469	84.66				
– Death	85	15.34				

3.3 Models and Inference Procedure

3.3.1 Model formulation and notation

Longitudinal measurements are considered as the realizations of a certain marker process at finite time points, and we use $Y(t)$ to denote the value of such a marker process at time t. We let T be survival time, and suppose that the survival time T is possibly right censored and the right-censoring time is missing at random. Suppose a set of n subjects are followed over an interval $[0, \tau]$, where τ is the study end time. Denote $\boldsymbol{b}_i$, $i = 1, \ldots, n$, as a vector of subject-specific random effects of dimension d_b and $\boldsymbol{b}_i$'s are mutually independent and identically distributed from a multivariate normal with mean zero and covariance matrix $\boldsymbol{\Sigma}_b$.

Given the random effects $\boldsymbol{b}_i$, the observed covariates, and the observed outcome history till time t, we assume that the longitudinal outcome $Y_i(t)$ at time t for subject

i follows a distribution from the exponential family with density

$$\exp\left\{\frac{y_i\eta_i(t) - B(\eta_i(t))}{A(D_i(t;\phi))} + C(y_i, D_i(t;\phi))\right\} \tag{3.1}$$

with $\mu_i(t) = \mathrm{E}(Y_i(t)|\boldsymbol{b}_i) = B'(\eta_i(t))$ and $v_i(t) = \mathrm{Var}(Y_i(t)|\boldsymbol{b}_i) = B''(\eta_i(t))A(D_i(t;\phi))$, satisfying

$$\eta_i(t) = g(\mu_i(t)) = \boldsymbol{X}_i(t)\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_i(t)\boldsymbol{b}_i$$

and $v_i(t) = v(\mu_i(t))A(D_i(t;\phi))$, where $g(\cdot)$ and $v(\cdot)$ are known link and variance functions respectively, and $\boldsymbol{X}_i(t)$ and $\widetilde{\boldsymbol{X}}_i(t)$ are the row vectors of the observed covariates for subject i, and $\boldsymbol{\beta}$ is a column vector of coefficients for $\boldsymbol{X}_i(t)$. The random effect $\boldsymbol{b}_i$ is allowed to differ for different individuals. Additionally, $\boldsymbol{X}_i(t)$ and $\widetilde{\boldsymbol{X}}_i(t)$ can be completely different or share some components, and may include dummy variables for different strata.

Given the random effects $\boldsymbol{b}_i$, the observed covariates, and the observed survival history before time t, the conditional hazard rate function for the survival time T_i of subject i is assumed to follow a stratified multiplicative hazards model

$$\lambda_s(t)\exp\{\widetilde{\boldsymbol{Z}}_i(t)(\boldsymbol{\psi}\circ\boldsymbol{b}_i) + \boldsymbol{Z}_i(t)\boldsymbol{\gamma}\}, \tag{3.2}$$

where $\boldsymbol{Z}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ are the row vectors of the observed covariates and may share some components, $\boldsymbol{\psi}$ is a vector of parameters of the coefficients for random effects, $\lambda_s(t)$ is the s-th stratum baseline hazard rate function, and $\boldsymbol{\gamma}$ is a column vector of coefficients for $\boldsymbol{Z}_i(t)$. Note that $\boldsymbol{Z}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ do not include dummy variables for strata since baseline hazard rate is stratum-specific. Here, for any vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ of the same dimension, $\boldsymbol{a}_1\circ\boldsymbol{a}_2$ denotes the component-wise product. In addition, $\widetilde{\boldsymbol{X}}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ have the same dimensions as $\boldsymbol{b}_i$'s.

Under models (3.1) and (3.2), the two outcomes $Y(t)$ and T are independent conditional on the covariates and random effect. The parameter $\boldsymbol{\psi}$ in model (3.2) characterizes the dependence between the longitudinal outcomes and the survival time due to latent random effect: $\boldsymbol{\psi} = \mathbf{0}$ means that the dependence between the survival time and longitudinal responses are not due to these latent variables; $\boldsymbol{\psi} \neq \mathbf{0}$ means that such dependence may be due to these latent variables. In other words, $\boldsymbol{\psi} > \mathbf{0}$ implies that there may be some latent factors increasing both the longitudinal outcomes and the risk of survival endpoint simultaneously while $\boldsymbol{\psi} < \mathbf{0}$ implies that some latent factors causing the increment of longitudinal outcomes may decrease the risk of survival endpoint.

We let n_i be the number of the observed longitudinal measurements for subject i, and assume that n_i and the observation times for longitudinal measurements are not informative about parameters of interest. That is, the distributions of n_i and the observation times for longitudinal measurements are independent of the parameters of interest in this joint model. The observed data from n subjects are $(n_i, Y_{ij}, \boldsymbol{X}_{ij}, \widetilde{\boldsymbol{X}}_{ij})$, $j = 1, \ldots, n_i,\ i = 1, \ldots, n,$ and $(V_i, \Delta_i, S_i, \{(\boldsymbol{Z}_i(t), \widetilde{\boldsymbol{Z}}_i(t)) : t \leq V_i\}),\ i = 1, \ldots, n,$ where for subject i, $(Y_{ij}, \boldsymbol{X}_{ij}, \widetilde{\boldsymbol{X}}_{ij})$ is the j-th observation of $(Y_i(t), \boldsymbol{X}_i(t), \widetilde{\boldsymbol{X}}_i(t))$, C_i is the right-censoring time, $V_i = \min(T_i, C_i)$, S_i denotes the stratum, and $\Delta_i = I(T_i \leq C_i)$.

Our goal is to estimate and make inferences on the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\phi}^T, \mathrm{Vec}(\boldsymbol{\Sigma}_b)^T, \boldsymbol{\psi}^T, \boldsymbol{\gamma}^T)^T$ and the baseline cumulative hazard functions with S strata, $\boldsymbol{\Lambda}(t) = (\Lambda_1(t), \ldots, \Lambda_S(t))^T$, where $\Lambda_s(t) = \int_0^t \lambda_s(u)du$, $s = 1, \ldots, S$. $\mathrm{Vec}(\cdot)$ operator creates a column vector from a matrix by stacking the diagonal and upper-triangle elements of the matrix.

3.3.2 Inference procedure

For all n subjects, we write $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T$, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, $\boldsymbol{V} = (V_1, \ldots, V_n)^T$, and $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T)^T$. Then, the likelihood function of the complete data $(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ for

$(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ has the form,

$$
\begin{aligned}
& L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) \\
&= \prod_{s=1}^{S} \prod_{i=1}^{n} \big[f(\boldsymbol{Y}_i, V_i | \boldsymbol{b}_i) f(\boldsymbol{b}_i) \big]^{I(S_i=s)} = \prod_{i=1}^{n} f(\boldsymbol{Y}_i | \boldsymbol{b}_i) \Bigg(\prod_{s=1}^{S} \big[f(V_i | \boldsymbol{b}_i) \big]^{I(S_i=s)} \Bigg) f(\boldsymbol{b}_i) \\
&= \prod_{i=1}^{n} \exp \left\{ \sum_{j=1}^{n_i} \left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i) - B(\boldsymbol{\beta}; \boldsymbol{b}_i)}{A(D_i(t_j;\phi))} + C(Y_{ij}; D_i(t_j;\phi)) \right] \right\} \\
&\quad \times \Bigg(\prod_{s=1}^{S} \Bigg[\lambda_s(V_i)^{\Delta_i} \exp \Bigg\{ \Delta_i \big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma} \big] \\
&\qquad\qquad - \int_0^{V_i} \exp \big\{ \widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma} \big\} d\Lambda_s(u) \Bigg\} \Bigg]^{I(S_i=s)} \Bigg) \\
&\quad \times (2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp \left\{ -\frac{1}{2} \boldsymbol{b}_i^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}_i \right\},
\end{aligned}
$$

and the full likelihood function of the observed data $(\boldsymbol{Y}, \boldsymbol{V})$ for the parameter $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ is expressed as

$$
L_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) = \int_{\boldsymbol{b}} L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) d\boldsymbol{b}. \tag{3.3}
$$

The proposed estimation method is to calculate the maximum likelihood estimates for $(\boldsymbol{\theta}, \boldsymbol{\Lambda}(t))$ over a set in which $\boldsymbol{\theta}$ is in a bounded set and $\Lambda_s(t)$ of $\boldsymbol{\Lambda}(t)$ belongs to a space consisting of all the increasing functions with $\Lambda_s(0) = 0$, $s = 1, \ldots, S$. We let each $\Lambda_s(t)$ of $\boldsymbol{\Lambda}(t)$, $s = 1, \ldots, S$, be an increasing and right-continuous step function with jumps only at the observed failure times belonging to stratum s.

Denote $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}})$ as the maximum likelihood estimator for $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$. We let $l_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) = \log\{L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})\}$ and $l_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) = \log\{L_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V})\}$, and denote $U_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ and $U_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V})$ as the gradient vectors of the corresponding log-likelihood functions respectively. The EM-algorithm is used for calculating the maximum likelihood estimates. In the EM-algorithm, $\boldsymbol{b}_i$ is considered as missing data for $i = 1, \ldots, n$. Therefore, the M-step solves the conditional score equations from complete

data given observations, where the conditional expectation can be evaluated in E-step. The procedure involves iterating between the following two steps until convergence is achieved: at the k-th iteration,

(1) E-step Calculate the conditional expectations of some known functions of $\boldsymbol{b}_i$, needed in the next M-step, for subject i with $S_i = s$ given observations and the current estimate $(\boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)})$. To do this, denote $q(\boldsymbol{b}_i)$ and $E[q(\boldsymbol{b}_i)|\boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)}]$ as a known function and its conditional expectation, respectively. By some algebra, $E[q(\boldsymbol{b}_i)|\boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)}]$ can be expressed in terms of a vector of new variables z_G following a multivariate Gaussian distribution with mean zero. The conditional expection is calculated using the Gauss-Hermite Quadrature numerical approximation with 20 quadrature points.

(2) M-step After differentiating the conditional expectation of complete data log-likelihood function given observations and the current estimate $(\boldsymbol{\theta}^{(k)}, \boldsymbol{\Lambda}^{(k)})$, the updated estimator $(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)})$ can be obtained as follows: $(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\phi}^{(k+1)})$ solves the conditional expectation of complete data log-likelihood score equation using one-step Newton-Raphson iteration,

$$\mathrm{E}\left[U_c\left(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\phi}^{(k+1)}|\boldsymbol{\theta}^{(k)}, \boldsymbol{\Lambda}^{(k)}\right)\right] = \mathbf{0},$$

where $U_c(\boldsymbol{\beta}, \boldsymbol{\phi}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ is the first partial derivative of the complete data log-likelihood $l_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\phi})$;

$$\boldsymbol{\Sigma}_b^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\mathrm{E}\left[\boldsymbol{b}_i\boldsymbol{b}_i^T|\boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)}\right]I(S_i = s);$$

$(\boldsymbol{\psi}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)})$ solves the partial likelihood score equation from the full data using one-step Newton-Raphson iteration,

$$\sum_{i=1}^{n}\sum_{s=1}^{S}\Delta_i\left\{\begin{pmatrix} E[(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{b}_i)|\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)}] \\ \boldsymbol{Z}_i \end{pmatrix} - \frac{\sum_{l:V_l\geq V_i}\begin{pmatrix} E[(\widetilde{\boldsymbol{Z}}_l^T(V_i)\circ\boldsymbol{b}_l)\exp\{\widetilde{\boldsymbol{Z}}_l(V_i)(\boldsymbol{\psi}\circ\boldsymbol{b}_l)+\boldsymbol{Z}_l(V_i)\boldsymbol{\gamma}|\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)}] \\ E[\boldsymbol{Z}_l(V_i)\exp\{\widetilde{\boldsymbol{Z}}_l(V_i)(\boldsymbol{\psi}\circ\boldsymbol{b}_l)+\boldsymbol{Z}_l(V_i)\boldsymbol{\gamma}\}|\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)}] \end{pmatrix} I(S_l=s)}{\sum_{l:V_l\geq V_i}E[\exp\{\widetilde{\boldsymbol{Z}}_l(V_i)(\boldsymbol{\psi}\circ\boldsymbol{b}_l)+\boldsymbol{Z}_l(V_i)\boldsymbol{\gamma}\}|\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)}]I(S_l=s)}\right\}I(S_i=s)$$

$$=\mathbf{0};$$

$\Lambda_s^{(k+1)}$ is obtained as an empirical function which has jumps only at the observed failure time,

$$\Lambda_s^{(k+1)}(t)=\sum_{i:V_i\leq t}\frac{\Delta_i I(S_i=s)}{\sum_{l:V_l\geq V_i}\mathrm{E}\left[\exp\left\{\widetilde{\boldsymbol{Z}}_l(V_i)(\boldsymbol{\psi}^{(k+1)}\circ\boldsymbol{b}_l)+\boldsymbol{Z}_l(V_i)\boldsymbol{\gamma}^{(k+1)}\right\}|\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)}\right]I(S_l=s)}.$$

The expressions of the conditional expectation and the conditional score equations calculated in the E- and M-steps for binary and Poisson longitudinal outcomes with survival time are given respectively in Sections 3.3.3.1 and 3.3.3.2.

The observed information matrix is adopted to obtain the variance estimate for $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}}(t))$. For the numerical calculation of the observed information matrix, we consider $\Lambda_s\{V_i\}$, the jump size of $\Lambda_s(t)$ at V_i belonging to stratum s for which $\Delta_i=1$, instead of $\lambda_s(V_i)$. That is, $\boldsymbol{\Lambda}\{\cdot\}=(\boldsymbol{\Lambda}_1^T\{\cdot\},\ldots,\boldsymbol{\Lambda}_S^T\{\cdot\})^T$ with $\boldsymbol{\Lambda}_s\{\cdot\}=(\Lambda\{T_{s1}\},\ldots,\Lambda\{T_{sm_s}\})^T$ for m_s failure times among n_s subjects $(0\leq m_s\leq n_s)$ of the s-th stratum, $s=1,\ldots,S$. Then, by the Louis (1982) formula,

$$
\begin{aligned}
\mathrm{I}(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}) \;=\; & \mathrm{E}_{b|Y,V}[B_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) | \boldsymbol{Y}, \boldsymbol{V}] \\
& - \mathrm{E}_{b|Y,V}[U_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) U_c^T(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) | \boldsymbol{Y}, \boldsymbol{V}] \\
& + \mathrm{E}_{b|Y,V}[U_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V})] \, \mathrm{E}_{b|Y,V}[U_c^T(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V})],
\end{aligned}
$$

where $B_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ is the negative of the second derivative matrix for the complete data log-likelihood $l_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$. The variance of $\sqrt{n}\ \widehat{\boldsymbol{\theta}}$ is asymptotically equal to the corresponding sub-matrix of the inverse of the calculated observed information matrix. The variance of $\widehat{\boldsymbol{\Lambda}}(t)$ is obtained using the estimated variances and covariances corresponding to $\boldsymbol{\Lambda}\{\cdot\}$ from the inverse of the observed information matrix where $T \leq t$ at the observed failures. In the EM-algorithm for variance estimation, we evaluate these conditional expectations only at the last iteration of the EM procedure for point estimation, where U_f is zero.

3.3.3 EM algorithm – examples

3.3.3.1 Binary longitudinal data and survival time

(1) E-step : For binary longitudinal outcomes and survival time, we calculate the conditional expectation of $q(\boldsymbol{b}_i)$ for subject i with $S_i = s$ given the observations and the current estimate $(\boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)})$ for some known function $q(\cdot)$. The conditional expectation denoted by $E\big[q(\boldsymbol{b}_i)|\boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)}\big]$ can be expressed as the following:
Given the current estimate $(\boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)})$,

$$
E\big[q(\boldsymbol{b}_i)|\boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)}\big] \;=\; \frac{\int_{\boldsymbol{z}_G} q\big(R(\boldsymbol{z}_G)\big) K(\boldsymbol{z}_G) \exp\{-\boldsymbol{z}_G^T \boldsymbol{z}_G\} d\boldsymbol{z}_G}{\int_{\boldsymbol{z}_G} K(\boldsymbol{z}_G) \exp\{-\boldsymbol{z}_G^T \boldsymbol{z}_G\} d\boldsymbol{z}_G}, \tag{3.4}
$$

where

$$R(\boldsymbol{z}_G) = \left(\boldsymbol{\Sigma}_b^{(k)}\right)^{\frac{1}{2}}\left[\sqrt{2}\boldsymbol{z}_G + \left(\boldsymbol{\Sigma}_b^{(k)}\right)^{\frac{1}{2}}\left(\sum_{j=1}^{n_i} y_{ij}\widetilde{\boldsymbol{X}}_{ij}^T + \Delta_i\left(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{\psi}^{(k)}\right)\right)\right],$$

$$K(\boldsymbol{z}_G) = \exp\left\{-\sum_{j=1}^{n_i}\log\left(1+e^{\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(k)}+\widetilde{\boldsymbol{X}}_{ij}R(\boldsymbol{z}_G)}\right) - \int_0^{V_i} e^{\widetilde{\boldsymbol{Z}}_i(u)\left(\boldsymbol{\psi}^{(k)}\circ R(\boldsymbol{z}_G)\right)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}^{(k)}} d\Lambda_s^{(k)}(u)\right\}, \tag{3.5}$$

$\left(\boldsymbol{\Sigma}_b^{(k)}\right)^{\frac{1}{2}}$ is an unique non-negative square root of $\boldsymbol{\Sigma}_b^{(k)}$ (i.e. $\left(\boldsymbol{\Sigma}_b^{(k)}\right)^{\frac{1}{2}} \times \left(\boldsymbol{\Sigma}_b^{(k)}\right)^{\frac{1}{2}} = \boldsymbol{\Sigma}_b^{(k)}$), and $\boldsymbol{z}_G$ follows a multivariate Gaussian distribution with mean zero.

(2) M-step : Since the parameter $\boldsymbol{\phi}$ is set to 1 for logistic distribution, we estimate only $\boldsymbol{\beta}$ in the longitudinal process. $\boldsymbol{\beta}^{(k+1)}$ solves the conditional expectation of complete data log-likelihood score equation, using one-step Newton-Raphson iteration,

$$\begin{aligned}
&\mathrm{E}\left[U_c\left(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\theta}^{(k)},\boldsymbol{\Lambda}^{(k)}\right)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij} - \sum_{s=1}^{S}\mathrm{E}\left[\frac{\exp\{\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(k+1)}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\}}{1+\exp\{\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(k+1)}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\}}\middle|\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)}\right] I(S_i=s)\right)\boldsymbol{X}_{ij}^T = \mathbf{0},
\end{aligned}$$

where $U_c(\boldsymbol{\beta};\boldsymbol{Y},\boldsymbol{V},\boldsymbol{b})$ is the first partial derivative of the complete data log-likelihood $l_c(\boldsymbol{\theta},\boldsymbol{\Lambda};\boldsymbol{Y},\boldsymbol{V},\boldsymbol{b})$ with respect to $\boldsymbol{\beta}$. $\boldsymbol{\Sigma}_b^{(k+1)}$, $(\boldsymbol{\psi}^{(k+1)},\boldsymbol{\gamma}^{(k+1)})$, and $\Lambda_s^{(k+1)}$ have the same expressions as in Section 3.3.2.

3.3.3.2 Poisson longitudinal data and survival time

(1) E-step : For Poisson longitudinal outcomes and survival time, given the current estimate $(\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)})$, the conditional expectation denoted by $E\left[q(\boldsymbol{b}_i)|\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)}\right]$ can be expressed as in (3.4) with $R(\boldsymbol{z}_G)$ defined as in (3.5),

$$K(\boldsymbol{z}_G) = \exp\left\{-\sum_{j=1}^{n_i} e^{\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(k)}+\widetilde{\boldsymbol{X}}_{ij}R(\boldsymbol{z}_G)} - \int_0^{V_i} e^{\widetilde{\boldsymbol{Z}}_i(u)\left(\boldsymbol{\psi}^{(k)}\circ R(\boldsymbol{z}_G)\right)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}^{(k)}} d\Lambda_s^{(k)}(u)\right\},$$

and $\boldsymbol{z}_G$ follows a multivariate Gaussian distribution with mean zero.

(2) M-step : Since the parameter $\boldsymbol{\phi}$ is set to 1 for Poisson distribution, we estimate only $\boldsymbol{\beta}$ in the longitudinal process. $\boldsymbol{\beta}^{(k+1)}$ solves the conditional expectation of complete data log-likelihood score equation, using one-step Newton-Raphson iteration,

$$\begin{aligned}
&\mathrm{E}\left[U_c\left(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\theta}^{(k)},\boldsymbol{\Lambda}^{(k)}\right)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij}-\sum_{s=1}^{S}\mathrm{E}\left[\exp\left\{\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(k+1)}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\right\}\middle|\boldsymbol{\theta}^{(k)},\Lambda_s^{(k)}\right]I(S_i=s)\right)\boldsymbol{X}_{ij}^T \;=\; \mathbf{0},
\end{aligned}$$

where $U_c(\boldsymbol{\beta};\boldsymbol{Y},\boldsymbol{V},\boldsymbol{b})$ is the first partial derivative of the complete data log-likelihood $l_c(\boldsymbol{\theta},\boldsymbol{\Lambda};\boldsymbol{Y},\boldsymbol{V},\boldsymbol{b})$ with respect to $\boldsymbol{\beta}$. $\boldsymbol{\Sigma}_b^{(k+1)}$, $(\boldsymbol{\psi}^{(k+1)},\boldsymbol{\gamma}^{(k+1)})$, and $\Lambda_s^{(k+1)}$ have the same expressions as in Section 3.3.2.

3.4 Asymptotic Properties

To study the asymptotic properties of the proposed estimator $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}}(t))$ with $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\phi}}^T,\mathrm{Vec}(\widehat{\boldsymbol{\Sigma}}_b)^T,\widehat{\boldsymbol{\psi}}^T,\widehat{\boldsymbol{\gamma}}^T)^T$ and $\widehat{\boldsymbol{\Lambda}}(t) = (\widehat{\Lambda}_1(t),\ldots,\widehat{\Lambda}_S(t))^T$, we assume the following conditions below.

(A1) The true parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T,\boldsymbol{\phi}_0^T,\mathrm{Vec}(\boldsymbol{\Sigma}_{b0})^T,\boldsymbol{\psi}_0^T,\boldsymbol{\gamma}_0^T)^T$ belongs to a known compact set Θ which lies in the interior of the domain for $\boldsymbol{\theta}$.

(A2) The true baseline hazard rate function $\boldsymbol{\lambda}_0(t) = (\lambda_{10}(t),\ldots,\lambda_{S0}(t))$ is bounded and positive in $[0,\tau]$, where τ is the time of study end.

(A3) For the censoring time C, $P(C \geq \tau|\boldsymbol{Z},\widetilde{\boldsymbol{Z}},\boldsymbol{X},\widetilde{\boldsymbol{X}}) = P(C = \tau|\boldsymbol{Z},\widetilde{\boldsymbol{Z}},\boldsymbol{X},\widetilde{\boldsymbol{X}}) > 0$.

(A4) For the number of observed longitudinal measurements per subject n_N, $P(n_N > d_b|\boldsymbol{X},\widetilde{\boldsymbol{X}}) > 0$ with probability one, and $P(n_N \leq n_0) = 1$ for some integer n_0.

(A5) Both $\boldsymbol{X}^T\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}$ are full rank with positive probability. Moreover, if there

exist constant vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ such that, with positive probability, for any t, $\boldsymbol{Z}(t)\boldsymbol{c}_1 = \alpha_0(t)$ and $\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{c}_2 = 0$ for a deterministic function $\alpha_0(t)$, then $\boldsymbol{c}_1 = 0$, $\boldsymbol{c}_2 = 0$, and $\alpha_0(t) = 0$.

Assumption (A3) means that, by the end of the study, some proportion of the subjects will still be alive and censored at the study end time τ, and thus the maximum right censoring time is equal to τ. Assumption (A4) implies that some proportion of the subjects have at least d_b longitudinal observations, and there exists an integer n_0 such that $P(n_N \leq n_0) = 1$. Consistency and asymptotic distribution of the proposed estimator are summarized in the following two theorems. We will present outlines of the proofs here. The detailed technical proofs are given in Section 3.5.

Theorem 3.1. *Under the assumptions (A1)~(A5), as $n \to \infty$, the maximum likelihood estimator $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}(t))$ is consistent under the product norm of the Euclidean distance and the supreme norm on $[0, \tau]$. That is, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} \|\widehat{\boldsymbol{\Lambda}}(t) - \boldsymbol{\Lambda}_0(t)\| \longrightarrow 0$, a.s., where $\|\widehat{\boldsymbol{\Lambda}}(t) - \boldsymbol{\Lambda}_0(t)\| = \sum_{s=1}^{S} |\widehat{\Lambda}_s(t) - \Lambda_{s0}(t)|$.*

Consistency of Theorem 3.1 can be proved by verifying the following three steps: First, we show that the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}})$ exists. This can be achieved by showing that the jump size $\Lambda_s\{V_i\}$, with $\Delta_i = 1$, is finite. Second, we show that, with probability one, $\widehat{\Lambda}_s(\tau)$, $s = 1, \ldots, S$, are bounded as $n \to \infty$. This can be proved by showing $\log \widehat{\Lambda}_s(\tau)$ is bounded. Third, given that the second step is true, by Helly's selection theorem (van der Vaart, 1998), we can choose a subsequence of $\widehat{\Lambda}_s(t)$ such that $\widehat{\Lambda}_s(t)$ weakly converges to some right-continuous monotone function $\Lambda_s^*(t)$ with probability one. For any sub-sequence, we can find a further sub-sequence, still denoted as $\widehat{\boldsymbol{\theta}}$, such that $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$. Using empirical process formulation and relevant Donsker properties with parameter identifiability, we can show that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda_s^* = \boldsymbol{\Lambda}_{s0}$, $s = 1, \ldots, S$. Based on these results, we can conclude that, with probability

one, $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\Lambda}}_s(t)$ converges to $\boldsymbol{\Lambda}_{s0}(t)$ in $[0,\tau]$, $s=1,\ldots,S$. Moreover, since $\boldsymbol{\Lambda}_{s0}(t)$ is right-continuous in $[0,\tau]$, the latter can be strengthened to uniform convergence; that is, $\sup_{t\in[0,\tau]}\|\widehat{\boldsymbol{\Lambda}}(t)-\boldsymbol{\Lambda}_0(t)\|\to 0$ almost surely.

Theorem 3.2. *Under the assumptions (A1)~(A5), as* $n\to\infty$, $\sqrt{n}\big((\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T,(\widehat{\boldsymbol{\Lambda}}(t)-\boldsymbol{\Lambda}_0(t))^T\big)^T$ *weakly converges to a Gaussian random element in* $R^{d_\theta}\times\ell^\infty[0,\tau]\times\cdots\times\ell^\infty[0,\tau]$, *and the estimator* $\widehat{\boldsymbol{\theta}}$ *is asymptotically efficient, where* d_θ *is the dimension of* $\boldsymbol{\theta}$ *and* $\ell^\infty[0,\tau]$ *is the normed space containing all the bounded functions in* $[0,\tau]$.

Once consistency is held, the conditions of Theorem 3.3.1 in van der Vaart and Wellner (1996), which implies the asymptotic normality of Theorem 3.2, are verified via the tools of empirical processes. These conditions are restated in Theorem 4 of Parner (1998). The smooth conditions in Theorem 4 of Parner (1998) can be verified using the regularity of the log-likelihood function in terms of model parameters and the Donsker properties of the score operators. In particular, in the invertibility condition of the information operator in Theorem 4 of Parner (1998), the verification of the one-to-one property of the information operator is specific to our proposed models and requires non-trivial work. Therefore, by Theorem 3.3.1 of van der Vaart and Wellner (1996), $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}_s-\Lambda_{s0})$ weakly converges to a Gaussian process, and by Proposition 3.3.1 in Bickel *et al.* (1993), $\widehat{\boldsymbol{\theta}}$ is an efficient estimator for $\boldsymbol{\theta}_0$.

3.5 Technical Details – Proofs for Asymptotic Properties

In this section, we present the detailed technical proofs for the asymptotic properties of the proposed estimator $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}}(t))$ with $\widehat{\boldsymbol{\theta}}=(\widehat{\boldsymbol{\beta}}^T,\widehat{\boldsymbol{\phi}}^T,\mathrm{Vec}(\widehat{\boldsymbol{\Sigma}}_b)^T,\widehat{\boldsymbol{\psi}}^T,\widehat{\boldsymbol{\gamma}}^T)^T$ and $\widehat{\boldsymbol{\Lambda}}(t)=(\widehat{\Lambda}_1(t),\ldots,\widehat{\Lambda}_S(t))^T$. Meanwhile, the supplementary proofs needed to prove the asymptotic properties are provided in Section 3.5.3. From the full likelihood function

of observed data $(\boldsymbol{Y}, \boldsymbol{V})$ for $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$,

$$
\begin{aligned}
L_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) &= \int_{\boldsymbol{b}} L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) d\boldsymbol{b} \\
&= \prod_{i=1}^{n} \int_{\boldsymbol{b}} \exp\left\{ \sum_{j=1}^{n_i} \left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i) - B(\boldsymbol{\beta}; \boldsymbol{b}_i)}{A(D_i(t_j;\phi))} + C(Y_{ij}; D_i(t_j;\phi)) \right] \right\} \\
&\quad \times \left(\prod_{s=1}^{S} \left[\lambda_s(V_i)^{\Delta_i} \exp\left\{ \Delta_i \big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma} \big] \right.\right.\right. \\
&\quad \left.\left.\left. - \int_0^{V_i} \exp\big\{ \widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma} \big\} d\Lambda_s(u) \right\} \right]^{I(S_i = s)} \right) \\
&\quad \times (2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\left\{ -\frac{1}{2} \boldsymbol{b}_i^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}_i \right\} d\boldsymbol{b},
\end{aligned}
$$

we have the observed log-likelihood function

$$
\begin{aligned}
&\sum_{i=1}^{n} \log \left[\int_{\boldsymbol{b}} \exp\left\{ \sum_{j=1}^{n_i} \left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i) - B(\boldsymbol{\beta}; \boldsymbol{b}_i)}{A(D_i(t_j;\phi))} + C(Y_{ij}; D_i(t_j;\phi)) \right] \right\} \right. \\
&\quad \times \left(\prod_{s=1}^{S} \left[\lambda_s(V_i)^{\Delta_i} \exp\left\{ \Delta_i \big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma} \big] \right.\right.\right. \\
&\quad \left.\left.\left. - \int_0^{V_i} \exp\big\{ \widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma} \big\} d\Lambda_s(u) \right\} \right]^{I(S_i = s)} \right) \\
&\quad \left. \times (2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\left\{ -\frac{1}{2} \boldsymbol{b}_i^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}_i \right\} d\boldsymbol{b} \right].
\end{aligned}
$$

Then, we obtain the following modified object function by replacing $\lambda_s(V_i)$ with $\Lambda_s\{V_i\}$ in the above expression where $\Lambda_s\{V_i\}$ is the jump size of $\Lambda_s(t)$ at the observed time V_i with $\Delta_i = 1$,

$$
\begin{aligned}
l_n(\boldsymbol{\theta}, \boldsymbol{\Lambda}) &= \sum_{i=1}^{n} \log \left[\int_{\boldsymbol{b}} \exp\left\{ \sum_{j=1}^{n_i} \left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i) - B(\boldsymbol{\beta}; \boldsymbol{b}_i)}{A(D_i(t_j;\phi))} + C(Y_{ij}; D_i(t_j;\phi)) \right] \right\} \right. \\
&\quad \times \left(\prod_{s=1}^{S} \left[\Lambda_s\{V_i\}^{\Delta_i} \exp\left\{ \Delta_i \big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma} \big] \right.\right.\right. \\
&\quad \left.\left.\left. - \int_0^{V_i} \exp\big\{ \widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma} \big\} d\Lambda_s(u) \right\} \right]^{I(S_i = s)} \right)
\end{aligned}
$$

$$\times (2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{b}_i^T\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}_i\right\}d\boldsymbol{b}\Bigg], \tag{3.6}$$

and $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}})$ maximizes $l_n(\boldsymbol{\theta},\boldsymbol{\Lambda})$ over the space $\{(\boldsymbol{\theta},\boldsymbol{\Lambda}) : \boldsymbol{\theta} \in \Theta, \boldsymbol{\Lambda} \in \mathbb{W}_n \times \mathbb{W}_n \cdots \times \mathbb{W}_n\}$, where $\mathbb{W}_n$ consists of all the right-continuous step functions only; that is, $\boldsymbol{\Lambda} = (\Lambda_1,\ldots,\Lambda_S)^T, s = 1,\ldots,S, \Lambda_s \in \mathbb{W}_n$. For the proofs of both Theorem 3.1 and Theorem 3.2, the modified object function is used in the place of the observed log-likelihood function.

3.5.1 Proof of consistency

Consistency can be proved by verifying the following three steps: First, we show the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}})$ exists. Second, we show that, with probability one, $\widehat{\Lambda}_s(\tau)$, $s = 1,\ldots,S$, are bounded as $n \to \infty$. Third, if the second step is true, by Helly's selection theorem (p9 of van der Vaart, 1998), we can choose a subsequence of $\widehat{\Lambda}_s$ such that $\widehat{\Lambda}_s$ weakly converges to some right-continuous monotone function Λ_s^* with probability one; that is, the measure given by $\mu_s([0,t]) = \widehat{\Lambda}_s(t)$ for $t \in [0,\tau]$ weakly converges to the measure given by $\mu_s^*([0,t]) = \Lambda_s^*(t)$. By choosing a sub-sequence, we can further assume $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$. Thus, in this third step, we show $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda_s^* = \boldsymbol{\theta}_{s0}$, $s = 1,\ldots,S$.

Once the three steps are completed, we can conclude that, with probability one, $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\Lambda}}_s$ converges to $\boldsymbol{\Lambda}_{s0}$ in $[0,\tau]$, $s = 1,\ldots,S$. However, since $\boldsymbol{\Lambda}_{s0}$ is continuous in $[0,\tau]$, the latter can be strengthened to uniform convergence; that is, $\sup_{t\in[0,\tau]}||\widehat{\boldsymbol{\Lambda}}(t)-\boldsymbol{\Lambda}_0(t)|| \to 0$ almost surely. Then, the proof of Theorem 3.1 will be done.

In the first step, we will show the existence of the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}})$. Since $\boldsymbol{\theta}$ belongs to a compact set Θ by the assumption (A1), it is sufficient to show that $\Lambda_s\{V_i\}$, the jump size of Λ_s at V_i for which $\Delta_i = 1$, is finite. Since, for each

subject i with $\Delta_i = 1$,

$$\begin{aligned}
&\Lambda_s\{V_i\}\exp\left\{\int_0^{V_i}\exp\left\{\widetilde{\boldsymbol{Z}}_i(t)(\boldsymbol{\psi}\circ\boldsymbol{b})+\boldsymbol{Z}_i(t)\boldsymbol{\gamma}\right\}d\Lambda_s(t)\right\}\\
&\qquad\leq\ \exp\left\{-2\big(\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi}\circ\boldsymbol{b})+\boldsymbol{Z}_i(V_i)\boldsymbol{\gamma}\big)\right\}\big(\Lambda_s\{V_i\}\big)^{-1},
\end{aligned}$$

we have that, from (3.6),

$$\begin{aligned}
l_n(\boldsymbol{\theta},\boldsymbol{\Lambda}) \ \leq\ & \sum_{i=1}^{n}\log\int_{\boldsymbol{b}}\left[\exp\left\{\sum_{j=1}^{n_i}\left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b})}{A(D_i(t_j;\phi))}+C(Y_{ij};D_i(t_j;\phi))\right]\right\}\right.\\
&\times\left(\prod_{s=1}^{S}\left[\big(\Lambda_s\{V_i\}\big)^{-\Delta_i}\exp\left\{-\Delta_i[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi}\circ\boldsymbol{b})+\boldsymbol{Z}_i(V_i)\boldsymbol{\gamma}]\right\}\right]^{I(S_i=s)}\right)\\
&\left.\times\,(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}\right\}\right]d\boldsymbol{b}.
\end{aligned}$$

Thus, if $\Lambda_s\{V_i\}\to\infty$ for some i with $\Delta_i = 1$, then $l_n(\boldsymbol{\theta},\boldsymbol{\Lambda})\to-\infty$, which is contradictory to that $l_n(\boldsymbol{\theta},\boldsymbol{\Lambda})$ is bounded. Therefore, we conclude that $\Lambda_s\{\cdot\}$, the jump size of Λ_s for stratum s, must be finite. By the conclusion and the assumption (A1), the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}})$ exists.

In the second step, we will show that $\widehat{\Lambda}_s(\tau)$ is bounded as n goes to infinity with probability one. We define $\widehat{\zeta}_s = \log\widehat{\Lambda}_s(\tau)$ and rescale $\widehat{\Lambda}_s$ by the factor $e^{\widehat{\zeta}_s}$. Then, we let $\widetilde{\Lambda}_s$ denote the rescaled function; that is, $\widetilde{\Lambda}_s(t) = \widehat{\Lambda}_s(t)/\widehat{\Lambda}_s(\tau) = \widehat{\Lambda}_s(t)e^{-\widehat{\zeta}_s}$. thus, $\widetilde{\Lambda}_s(\tau) = 1$. To prove this second step, it is sufficient to show $\widehat{\zeta}_s$ is bounded. After some algebra in (3.6), we obtain that, for any $\boldsymbol{\Lambda}\in\mathbb{W}\times\mathbb{W}\cdots\times\mathbb{W}$,

$$\begin{aligned}
n^{-1}l_n(\widehat{\boldsymbol{\theta}},\boldsymbol{\Lambda}) \ =\ & \frac{1}{2}\sum_{i=1}^{n}\left[\sum_{j=1}^{n_i}\left(\frac{Y_{ij}\boldsymbol{X}_{ij}\widehat{\boldsymbol{\beta}}}{A(D_i(t_j;\widehat{\phi}))}+C(Y_{ij};D_i(t_j;\widehat{\phi}))\right)+\sum_{s=1}^{S}\Delta_i\big(\boldsymbol{Z}_i(V_i)\widehat{\boldsymbol{\gamma}}\big)I(S_i=s)\right.\\
&-\frac{1}{2}\log\left\{(2\pi)^{d_b}|\widehat{\boldsymbol{\Sigma}}_b|\right\}+\frac{1}{2}\boldsymbol{M}_i^T\boldsymbol{M}_i-\frac{1}{2}\log|\widehat{\boldsymbol{\Sigma}}_b|+\sum_{s=1}^{S}\Delta_i I(S_i=s)\log\Lambda_s\{V_i\}\\
&\left.+\log\int_{\boldsymbol{b}_0}\left[\exp\left\{-\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))}\right.\right.\right.
\end{aligned}$$

$$- \sum_{s=1}^{S} I(S_i = s) \int_0^{V_i} e^{Q_{1i}(t, \boldsymbol{b}_0, \widehat{\boldsymbol{\theta}})} d\Lambda_s(t) \Bigg\} \Bigg] d\boldsymbol{b}_0 \Bigg],$$

where
$$\begin{aligned} \boldsymbol{M}_i &= \widehat{\boldsymbol{\Sigma}}_b^{1/2} \Bigg(\sum_{j=1}^{n_i} \frac{Y_{ij} \widetilde{\boldsymbol{X}}_{ij}}{A(D_i(t_j; \widehat{\phi}))} + \sum_{s=1}^{S} I(S_i = s) \Delta_i \big(\widetilde{\boldsymbol{Z}}_i(V_i) \circ \widehat{\boldsymbol{\psi}}^T \big) \Bigg)^T, \\ \boldsymbol{b}_0 &= \boldsymbol{\Sigma}_b^{-1/2} \boldsymbol{b} - \boldsymbol{M}_i, \end{aligned}$$

and
$$Q_{1i}(t, \boldsymbol{b}_0, \widehat{\boldsymbol{\theta}}) = \big(\widetilde{\boldsymbol{Z}}_i(t) \circ \widehat{\psi}^T \big) \widehat{\boldsymbol{\Sigma}}_b^{1/2} \boldsymbol{b}_0 + \boldsymbol{Z}_i(t) \widehat{\boldsymbol{\gamma}} + \big(\widetilde{\boldsymbol{Z}}_i(t) \circ \widehat{\psi}^T \big) \widehat{\boldsymbol{\Sigma}}_b^{1/2} \boldsymbol{M}_i.$$

Thus, since $0 \leq n^{-1} l_n(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}) - n^{-1} l_n(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\Lambda}})$ where $\widehat{\boldsymbol{\Lambda}} = e^{\widehat{\boldsymbol{\xi}}} \circ \widetilde{\boldsymbol{\Lambda}}$, it follows that

$$\begin{aligned} 0 \leq\ & \frac{1}{n} \sum_{i=1}^{n} \sum_{s=1}^{S} \Delta_i I(S_i = s) \Bigg(\log e^{\widehat{\zeta}_s} \widetilde{\Lambda}_s - \log \widetilde{\Lambda}_s \Bigg) \\ & + \frac{1}{n} \sum_{i=1}^{n} \log \int_{\boldsymbol{b}_0} \Bigg[\exp \Bigg\{ - \frac{1}{2} \boldsymbol{b}_0^T \boldsymbol{b}_0 - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_0)}{A(D_i(t_j; \widehat{\phi}))} \\ & \qquad - \sum_{s=1}^{S} I(S_i = s) e^{\widehat{\zeta}_s} \int_0^{V_i} e^{Q_{1i}(t, \boldsymbol{b}_0, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Bigg\} \Bigg] d\boldsymbol{b}_0 \\ & - \frac{1}{n} \sum_{i=1}^{n} \log \int_{\boldsymbol{b}_0} \Bigg[\exp \Bigg\{ - \frac{1}{2} \boldsymbol{b}_0^T \boldsymbol{b}_0 - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_0)}{A(D_i(t_j; \widehat{\phi}))} \\ & \qquad - \sum_{s=1}^{S} I(S_i = s) \int_0^{V_i} e^{Q_{1i}(t, \boldsymbol{b}_0, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Bigg\} \Bigg] d\boldsymbol{b}_0. \quad (3.7) \end{aligned}$$

According to the assumption (A2), there exist some positive constants $C1$, $C2$ and $C3$ such that $|Q_{1i}(t, \boldsymbol{b}_0, \widehat{\boldsymbol{\theta}})| \leq C_1 \|\boldsymbol{b}_0\| + C_2 \|\boldsymbol{Y}_i\| + C_3$. By denoting $\boldsymbol{b}_0$ as a vector of variables following a standard multivariate normal distribution, from concavity of the logarithm function, in the third term of (3.7),

$$\begin{aligned} & \log \int_{\boldsymbol{b}_0} \Bigg[\exp \Bigg\{ - \frac{1}{2} \boldsymbol{b}_0^T \boldsymbol{b}_0 - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_0)}{A(D_i(t_j; \widehat{\phi}))} - \sum_{s=1}^{S} I(S_i = s) \int_0^{V_i} e^{Q_{1i}(t, \boldsymbol{b}_0, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Bigg\} \Bigg] d\boldsymbol{b}_0 \\ &= (2\pi)^{d_b/2} \log \mathrm{E}_{\boldsymbol{b}_0} \Bigg[\exp \Bigg\{ - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_0)}{A(D_i(t_j; \widehat{\phi}))} - \sum_{s=1}^{S} I(S_i = s) \int_0^{V_i} e^{Q_{1i}(t, \boldsymbol{b}_0, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Bigg\} \Bigg] \\ &\geq (2\pi)^{d_b/2} \log \mathrm{E}_{\boldsymbol{b}_0} \Bigg[\exp \Bigg\{ - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_0)}{A(D_i(t_j; \widehat{\phi}))} - e^{C_1 \|\boldsymbol{b}_0\| + C_2 \|\boldsymbol{Y}_i\| + C_3} \Bigg\} \Bigg] \end{aligned}$$

$$
\begin{aligned}
&\geq (2\pi)^{d_b/2}\,\mathrm{E}_{\boldsymbol{b}_0}\left[-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))}-e^{C_1\|\boldsymbol{b}_0\|+C_2\|\boldsymbol{Y}_i\|+C_3}\right]\\
&= -e^{C_2\|\boldsymbol{Y}_i\|+C_4}-C_5,
\end{aligned}
$$

where C_4 and C_5 are positive constants. Then, since it is easily verified that $\mathrm{E}_{\boldsymbol{b}_0}\left[\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))}+e^{C_1\|\boldsymbol{b}_0\|+C_2\|\boldsymbol{Y}_i\|+C_3}\right]<\infty$, by the strong law of large numbers and the assumption (A4), the third term of (3.7)

$$
\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\log\int_{\boldsymbol{b}_0}\left[\exp\left\{-\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))}-\sum_{s=1}^{S}I(S_i=s)\int_0^{V_i}e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\right\}\right]d\boldsymbol{b}_0\\
&\quad\leq \frac{1}{n}\sum_{i=1}^{n}(e^{C_2\|\boldsymbol{Y}_i\|+C_4}+C_5) \triangleq C_6
\end{aligned}
$$

can be bounded by some constant C_6 from above. Then (3.7) becomes

$$
\begin{aligned}
0 \leq\ & \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\Delta_i I(S_i=s)\widehat{\zeta}_s\\
&+\frac{1}{n}\sum_{i=1}^{n}\log\int_{\boldsymbol{b}_0}\left[\exp\left\{-\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))}\right.\right.\\
&\qquad\left.\left.-\sum_{s=1}^{S}e^{\widehat{\zeta}_s}\int_0^{V_i}e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\right\}\right]d\boldsymbol{b}_0+C_6\\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\Delta_i I(S_i=s)\widehat{\zeta}_s\\
&+\frac{1}{n}\sum_{i=1}^{n}I(V_i=\tau)\log\int_{\boldsymbol{b}_0}\left[\exp\left\{-\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))}\right.\right.\\
&\qquad\left.\left.-\sum_{s=1}^{S}e^{\widehat{\zeta}_s}\int_0^{\tau}e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\right\}\right]d\boldsymbol{b}_0\\
&+\frac{1}{n}\sum_{i=1}^{n}I(V_i\neq\tau)\log\int_{\boldsymbol{b}_0}\left[\exp\left\{-\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))}\right\}\right]d\boldsymbol{b}_0+C_6\\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\Delta_i I(S_i=s)\widehat{\zeta}_s\\
&+\frac{1}{n}\sum_{i=1}^{n}I(V_i=\tau)\log\int_{\boldsymbol{b}_0}\left[\exp\left\{-\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))}\right.\right.
\end{aligned}
$$

$$- \sum_{s=1}^{S} e^{\widehat{\zeta}_s} \int_0^\tau e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Big\} \Big] d\boldsymbol{b}_0 + C_7, \tag{3.8}$$

where C_7 is a constant. On the other hand, since, for any $\Gamma \geq 0$ and $x > 0$, $\Gamma \log\big(1 + x/\Gamma\big) \leq \Gamma x/\Gamma = x$, we have that $e^{-x} \leq \big(1 + x/\Gamma\big)^{-\Gamma}$. Therefore, in the second term of (3.8),

$$\begin{aligned}
&\exp\Big\{ -\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0 - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))} - \sum_{s=1}^{S} e^{\widehat{\zeta}_s} \int_0^\tau e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Big\} \\
&\leq \exp\Big\{ -\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0 - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))} \Big\} \times \Big\{ 1 + \frac{\sum_{s=1}^{S} I(S_i = s) e^{\widehat{\zeta}_s} \int_0^\tau e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t)}{\Gamma} \Big\}^{-\Gamma} \\
&\leq \Gamma^\Gamma \times \exp\Big\{ -\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0 - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))} \Big\} \times \Big\{ \sum_{s=1}^{S} I(S_i = s) e^{\widehat{\zeta}_s} \int_0^\tau e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Big\}^{-\Gamma} \\
&= \Gamma^\Gamma \times \exp\Big\{ -\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0 - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))} - \Gamma \sum_{s=1}^{S} I(S_i = s) e^{\widehat{\zeta}_s} \Big\} \\
&\qquad \times \Big\{ \int_0^\tau e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Big\}^{-\Gamma}.
\end{aligned} \tag{3.9}$$

Since $Q_{1i}(t, \boldsymbol{b}_0, \widehat{\boldsymbol{\theta}}) \geq -C_1\|\boldsymbol{b}_0\| - C_2\|\boldsymbol{Y}_i\| - C_3$,

$$\begin{aligned}
\int_0^\tau e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) &\geq \int_0^\tau e^{-C_1\|\boldsymbol{b}_0\| - C_2\|\boldsymbol{Y}_i\| - C_3} d\widetilde{\Lambda}_s(t) \\
&= e^{-C_1\|\boldsymbol{b}_0\| - C_2\|\boldsymbol{Y}_i\| - C_3} \times \big\{ \widetilde{\Lambda}_s(\tau) - \widetilde{\Lambda}_s(0) \big\} \\
&= e^{-C_1\|\boldsymbol{b}_0\| - C_2\|\boldsymbol{Y}_i\| - C_3}.
\end{aligned}$$

Thus, in (3.9), $$\Big\{ \int_0^\tau e^{Q_{1i}(t,\boldsymbol{b}_0,\widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t) \Big\}^{-\Gamma} \leq e^{C_1\Gamma\|\boldsymbol{b}_0\| + C_2\Gamma\|\boldsymbol{Y}_i\| + C_3\Gamma},$$

and $$\begin{aligned}
(3.9) \leq \int_{\boldsymbol{b}_0} \Big[\Gamma^\Gamma \times \exp\Big\{ &-\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0 - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))} - \Gamma \sum_{s=1}^{S} I(S_i = s) e^{\widehat{\zeta}_s} \\
&+ C_1\Gamma\|\boldsymbol{b}_0\| + C_2\Gamma\|\boldsymbol{Y}_i\| + C_3\Gamma \Big\} \Big] d\boldsymbol{b}_0.
\end{aligned}$$

Therefore, (3.8) gives that

$$\begin{aligned} 0 \;\le\; & C_7 + \frac{1}{n}\sum_{i=1}^{n}\Delta_i\Big(\sum_{s=1}^{S}\widehat{\zeta}_s\Big) + \frac{1}{n}\sum_{i=1}^{n} I(V_i=\tau)\log\Big\{\Gamma^{\Gamma}\times\exp\Big\{-\Gamma\sum_{s=1}^{S} I(S_i=s)\widehat{\zeta}_s\Big\} \\ & \times \int_{\boldsymbol{b}_0}\Big[\exp\Big\{-\frac{1}{2}\boldsymbol{b}_0^T\boldsymbol{b}_0 - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_0)}{A(D_i(t_j;\widehat{\phi}))} + C_1\Gamma\|\boldsymbol{b}_0\| + C_2\Gamma\|\boldsymbol{Y}_i\| + C_3\Gamma\Big\}\Big]d\boldsymbol{b}_0\Big\} \\ =\; & C_7 + \frac{1}{n}\sum_{i=1}^{n}\Delta_i\Big(\sum_{s=1}^{S}\widehat{\zeta}_s\Big) - \frac{\Gamma}{n}\sum_{i=1}^{n} I(V_i=\tau)\Big(\sum_{s=1}^{S}\widehat{\zeta}_s\Big) + C_8(\Gamma), \end{aligned} \tag{3.10}$$

where $C_8(\Gamma)$ is a deterministic function of Γ. For the s-th stratum, (3.10) is that

$$0 \;\le\; C_7 + \sum_{i=1}^{n}\Delta_i I(S_i=s)\widehat{\zeta}_s - \frac{\Gamma}{n}\sum_{i=1}^{n} I(V_i=\tau)I(S_i=s)\widehat{\zeta}_s + C_8(\Gamma).$$

By the strong law of large numbers, $\sum_{i=1}^{n} I(V_i=\tau)I(S_i=s)/n \longrightarrow P(V_i=\tau,S_i=s)>0$. Then, we can choose Γ large enough such that $\sum_{i=1}^{n}\Delta_i I(S_i=s)/n \;\le\; (\Gamma/2n)\sum_{i=1}^{n} I(V_i=\tau)I(S_i=s)$. Thus, we obtain that

$$0 \;\le\; C_7 + C_8(\Gamma) - \frac{\Gamma}{2n}\sum_{i=1}^{n} I(V_i=\tau)I(S_i=s)\widehat{\zeta}_s.$$

In other words,

$$\widehat{\zeta}_s \;\le\; \frac{(C_7+C_8(\Gamma))2n}{\Gamma\sum_{i=1}^{n} I(V_i=\tau)I(S_i=s)} \;\longrightarrow\; \frac{(C_7+C_8(\Gamma))2}{\Gamma P(V_i=\tau,S_i=s)}.$$

If we denote $B_{s0}=\exp\big\{2\big(C_7+C_8(\Gamma)\big)/\big(\Gamma P(V_i=\tau,S_i=s)\big)\big\}$, we conclude that $\widehat{\Lambda}_s(\tau)\le B_{s0}, s=1,\ldots,S$. Note that the above arguments hold for every sample in the probability space except a set with zero probability. Therefore, we have shown that, with probability one, $\widehat{\Lambda}_s(\tau)$ is bounded for any sample size n.

In the third step, the goal of this step is to show that, if $\widehat{\boldsymbol{\theta}}\to\boldsymbol{\theta}^*$ and $\widehat{\Lambda}_s$ weakly converges to Λ^* with probability one, then $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$ and $\Lambda_s^*=\boldsymbol{\theta}_{s0}$, $s=1,\ldots,S$. We set some preliminaries as the followings: For convenience, we omit the index i for

subject and use $\boldsymbol{O}$ to abbreviate the observed statistics $(\boldsymbol{Y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}, V, \Delta, n_N, s)$ and $\{\boldsymbol{Z}(t), \widetilde{\boldsymbol{Z}}(t), 0 \le t \le V\}$ for a subject. By dropping $(\lambda_s(V))^{\Delta}$ from the complete data likelihood function, we define that

$$
\begin{aligned}
G(\boldsymbol{b}, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) &= \exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_i\boldsymbol{b}) - B(\boldsymbol{\beta}; \boldsymbol{b})}{A(D(t_j;\phi))} + C(Y_j; D(t_j;\phi))\right]\right\} \\
&\quad \times \exp\left\{\Delta_i\big[\widetilde{\boldsymbol{Z}}(V)(\boldsymbol{\psi} \circ \boldsymbol{b}) + \boldsymbol{Z}(V)\boldsymbol{\gamma}\big]\right. \\
&\qquad \left. - \int_0^V \exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi} \circ \boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}\big\} d\Lambda_s(t)\right\} \\
&\quad \times (2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}\right\},
\end{aligned}
$$

$$
\text{and} \qquad Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) = \frac{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) \exp\big\{\widetilde{\boldsymbol{Z}}(v)(\boldsymbol{\psi} \circ \boldsymbol{b}) + \boldsymbol{Z}(v)\boldsymbol{\gamma} d\boldsymbol{b}\big\}}{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) d\boldsymbol{b}}.
$$

Furthermore, for any measurable function $f(\boldsymbol{O})$, we use operator notation to define $\mathbf{P}_n f = n^{-1}\sum_{i=1}^n f(\boldsymbol{O}_i)$ and $\mathbf{P} f = \int f d\,\mathbf{P} = \mathrm{E}[f(\boldsymbol{O})]$. Thus, $\mathbf{P}_n f$ is the empirical measure from n i.i.d observations and $\sqrt{n}(\mathbf{P}_n - \mathbf{P})$ is the empirical process based on these observations. We also define a class $\mathscr{F} = \{Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) : v \in [0, \tau], \boldsymbol{\theta} \in \Theta, \Lambda_s \in \mathbb{W}, \Lambda_s(0) = 0, \Lambda_s(\tau) \le B_{s0}\}$, where B_{s0} is the constant given in the second step and $\mathbb{W}$ contains all nondecreasing functions in $[0, \tau]$. According to the result proved in Section 3.5.3.1, $\mathscr{F}$ is P-Donsker.

Let m_s denote the number of subjects in stratum s; i.e. $n = \sum_{s=1}^S m_s$. V_s and Δ_s denote the observed time and censoring indicator for a subject belonging to stratum s, respectively. Thus, V_{sk} and Δ_{sk} are the k-th subject observed time and censoring indicator in stratum s.

Now we start the proof of the third step. Since $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}})$ maximizes the function $l_n(\boldsymbol{\theta}, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_S)^T$ and Λ_s, $s = 1, \ldots, S$, are any step functions with

jumps only at V_i belonging to stratum s for which $\Delta_i = 1$, we differentiate $l_n(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ with respect to $\Lambda_s\{V_{sk}\}$ and obtain the following equation, satisfied by $\widehat{\Lambda}_s$,

$$\widehat{\Lambda}_s\{V_{sk}\} = \frac{\Delta_{sk}}{m_s \mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)\right\}\big|_{v=V_{sk}}}$$

Imitating the above equation, we also can construct another function, denoted by $\bar{\boldsymbol{\Lambda}} = (\bar{\Lambda}_1, \ldots, \bar{\Lambda}_S)^T$ such that $\bar{\Lambda}_s$, $s = 1, \ldots, S$, are also step functions with jumps only at the the observed V_{sk} and the jump size $\bar{\Lambda}_s\{V_{sk}\}$ is given by

$$\bar{\Lambda}_s\{V_{sk}\} = \frac{\Delta_{sk}}{m_s \mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_{sk}}}.$$

Equivalently,

$$\bar{\Lambda}_s(t) = \frac{1}{m_s}\sum_{k=1}^{m_s} \frac{I(V_{sk} \leq t)\Delta_{sk}}{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_{sk}}}.$$

Then, we claim $\bar{\Lambda}_s(t)$ uniformly converges to $\Lambda_{s0}(t)$ in $[0, \tau]$. To prove the claim, note

$$\begin{aligned}
&\sup_{t\in[0,\tau]}\left|\bar{\Lambda}_s(t) - \mathrm{E}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right]\right| \\
= &\sup_{t\in[0,\tau]}\left|\frac{1}{m_s}\sum_{k=1}^{m_s} \frac{I(V_{sk} \leq t)\Delta_{sk}}{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_{sk}}}\right. \\
&\qquad - \mathbf{P}_{m_s}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right] \\
&\quad + \mathbf{P}_{m_s}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right] \\
&\qquad \left. - \mathbf{P}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right]\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{t\in[0,\tau]}\left|\frac{1}{m_s}\sum_{k=1}^{m_s} I(V_{sk}\leq t)\Delta_{sk}\left[\frac{1}{\mathbf{P}_{m_s}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}}\right.\right. \\
&\qquad\qquad \left.\left. - \frac{1}{\mathbf{P}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}}\right]\right|_{v=V_{sk}}\Bigg| \\
&\quad + \sup_{t\in[0,\tau]}\left|\left(\mathbf{P}_{m_s}-\mathbf{P}\right)\left[\frac{I(V_s\leq t)\Delta_s}{\mathbf{P}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}\big|_{v=V_s}}\right]\right| \\
&\leq \sup_{t\in[0,\tau]}\left|\frac{1}{\mathbf{P}_{m_s}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}} - \frac{1}{\mathbf{P}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}}\right| \\
&\quad + \sup_{t\in[0,\tau]}\left|\left(\mathbf{P}_{m_s}-\mathbf{P}\right)\left[\frac{I(V_s\leq t)\Delta_s}{\mathbf{P}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}\big|_{v=V_s}}\right]\right|. \qquad (3.11)
\end{aligned}
$$

In (3.11), the right hand side converges to 0 because the first and second terms on the right hand side converges to 0 in the following: First, according to Section 3.5.3.1, $\{Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) : v \in [0,\tau]\}$ is a bounded and Glivenko-Cantelli class. $\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) : v \in [0,\tau]\}$ is also a Glivenko-Cantelli class because $\{I(V_s \geq v) : v \in [0,\tau]\}$ is a Glivenko-Cantelli class and the functional $(f,g) \to fg$ for any bounded two functions f and g is Lipschitz continuous. Then, we obtain that $\sup_{t\in[0,\tau]}\left|\mathbf{P}_{m_s}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\} - \mathbf{P}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}\right|$ converges to 0. Besides, from Section 3.5.3.1, $\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\} > \mathbf{P}\left\{I(V_s \geq v)\exp\{-C_9 - C_{10}\|\boldsymbol{Y}\|\}\right\}$ for the two constants C_9 and C_{10}, which means $\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}$ is bounded from below. Thus, the first term tends to 0. Second, since the class $\left\{I(V_s \leq t)\Delta_s / \mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}\big|_{v=V_s} : t \in [0,\tau]\right\}$ is also a Glivenko-Cantelli class, the second term vanishes as m_s goes to infinity.

Therefore, we conclude that $\bar{\Lambda}_s(t)$ uniformly converges to

$$
\mathrm{E}\left[\frac{I(V_s\leq t)\Delta_s}{\mathbf{P}\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}\big|_{v=V_s}}\right]. \qquad (3.12)
$$

We can easily verify that (3.12) is equal to $\Lambda_{s0}(t)$. Thus, the claim that $\bar{\Lambda}_s(t)$ uniformly converges to $\Lambda_{s0}(t)$ in $[0,\tau]$ has been proved.

From the construction of $\bar{\Lambda}_s(t)$, we obtain that

$$\widehat{\Lambda}_s(t) = \int_0^t \frac{d\widehat{\Lambda}_s(v)}{d\bar{\Lambda}_s(v)} d\bar{\Lambda}_s(v) = \int_0^t \frac{\mathbf{P}_{m_s}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}}{\mathbf{P}_{m_s}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)\right\}} d\bar{\Lambda}_s(v). \tag{3.13}$$

$\widehat{\Lambda}_s(t)$ is absolutely continuous with respect to $\bar{\Lambda}_s(t)$. On the other hand, since both $\{I(V_s \geq v) : v \in [0, \tau]\}$ and $\mathscr{F}$ are Glivenko-Cantelli classes, $\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) : v \in [0, \tau]\}$ is also a Glivenko-Cantelli class. Thus, we have

$$\begin{aligned} &\sup_{v \in [0,\tau]} \left|(\mathbf{P}_{m_s} - \mathbf{P})\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)\right\}\right| + \sup_{v \in [0,\tau]} \left|(\mathbf{P}_{m_s} - \mathbf{P})\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\right| \\ &\longrightarrow 0 \quad \text{a.s.} \end{aligned}$$

By the bounded convergence theorem and the fact that $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ and $\widehat{\Lambda}_s$ converges to Λ_s^*, for each v, $\mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)\right\} \longrightarrow \mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)\right\}$; moreover, it is straightforward to check the derivative of $\mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)\right\}$ with respect to v. Thus, by the Arzela-Ascoli theorem, uniformly in $[0, \tau]$,

$$\mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)\right\} \longrightarrow \mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)\right\}.$$

Then, combining the above result and (3.13), it holds that, uniformly in $[0, \tau]$,

$$\frac{\widehat{\Lambda}_s\{v\}}{\bar{\Lambda}_s\{v\}} = \frac{\mathbf{P}_{m_s}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}}{\mathbf{P}_{m_s}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)\right\}} \longrightarrow \frac{\mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}}{\mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)\right\}}. \tag{3.14}$$

After taking limits on both sides of (3.13), we obtain that

$$\Lambda_s^*(t) = \int_0^t \frac{\mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}}{\mathbf{P}\left\{I(V_s \geq v) Q(v, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)\right\}} d\Lambda_{s0}(v), \tag{3.15}$$

Therefore, since $\Lambda_{s0}(t)$ is differentiable with respect to the Lebesque measure, so is

$\Lambda_s^*(t)$; that is, (3.15) is equal to

$$\int_0^t \frac{d\Lambda_s^*(v)}{d\Lambda_{s0}(v)} d\Lambda_{s0}(v). \tag{3.16}$$

And we denote $\lambda_s^*(t)$ as the derivative of $\Lambda_s^*(t)$. Additionally, from (3.14) ~ (3.16), note that $\widehat{\Lambda}_s\{V_s\}/\bar{\Lambda}_s\{V_s\}$ uniformaly converges to $d\Lambda_s^*(V_s)/d\Lambda_{s0}(V_s) = \lambda_s^*(V_s)/\lambda_{s0}(V_s)$. Therefore, a second conclusion is that $\widehat{\Lambda}_s$ uniformly converges to Λ_s^* since Λ_s^* is continuous.

On the other hand,

$$\begin{aligned}
& n^{-1} l_n(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}) - n^{-1} l_n(\boldsymbol{\theta}_0, \bar{\boldsymbol{\Lambda}}) \\
&= \sum_{s=1}^{S} \left(\mathbf{P}_{m_s} \left[\Delta_s \log \frac{\widehat{\Lambda}_s\{V_s\}}{\Lambda_s\{V_s\}} \right] + \mathbf{P}_{m_s} \left[\log \frac{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s) d\boldsymbol{b}}{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \bar{\Lambda}_s) d\boldsymbol{b}} \right] \right) \\
&\geq 0.
\end{aligned} \tag{3.17}$$

Using the result of Section 3.5.3.1 and similar arguments as above, we can verify that

$$\log \frac{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s) d\boldsymbol{b}}{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \bar{\Lambda}_s) d\boldsymbol{b}}$$

belongs to a Glivenko-Cantelli class and

$$\mathbf{P}\left[\log \frac{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s) d\boldsymbol{b}}{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \bar{\Lambda}_s) d\boldsymbol{b}} \right] \longrightarrow \mathbf{P}\left[\log \frac{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}}{\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}} \right].$$

Since $\widehat{\Lambda}_s\{V_s\}/\bar{\Lambda}_s\{V_s\}$ uniformaly converges to $\lambda_s^*\{V_s\}/\lambda_{s0}\{V_s\}$, we obtain that, from (3.17),

$$\mathbf{P}\left[\log \left\{ \frac{\left(\lambda_s^*(V_s)\right)^{\Delta_s} \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}}{\left(\lambda_{s0}(V_s)\right)^{\Delta_s} \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}} \right\} \right] \geq 0.$$

Note that the left-hand side of the inequality is the negative Kullback-Leibler informa-

tion. Then, the equality holds with probability one, and it immediately follows

$$\left(\lambda_s^*(V_s)\right)^{\Delta_s} \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b} = \left(\lambda_{s0}(V_s)\right)^{\Delta_s} \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}. \tag{3.18}$$

Our proof will be completed if we can show $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda_s^* = \Lambda_{s0}$ from (3.18). Since (3.18) holds with probability one, (3.18) holds for any $(V_s, \Delta_s = 1)$ and the case $(V_s = \tau, \Delta_s = 0)$, but may not hold for $(V_s, \Delta_s = 0)$ when $V \in (0, \tau)$. However, we can show that (3.18) is also true for $(V_s, \Delta_s = 0)$ when $V_s \in (0, \tau)$. To do this, treating both sides of (3.18) as functions of V_s, we integrate these functions over an interval (V_s, τ) for $\Delta_s = 0$ as the following;

$$\int_{V_s}^{\tau} \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b} \quad = \quad \int_{V_s}^{\tau} \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}$$

to obtain that

$$\begin{aligned} &\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}\Big|_{\Delta_s=0, V_s=\tau} - \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}\Big|_{\Delta_s=0, V_s=V_s} \\ &\quad = \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}\Big|_{\Delta_s=0, V_s=\tau} - \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}\Big|_{\Delta_s=0, V_s=V_s}. \end{aligned}$$

After comparing this above equality with another following equality, which is given by (3.18) at $\Delta_s = 0$ and $V_s = \tau$,

$$\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}\Big|_{\Delta_s=0, V_s=\tau} \quad = \quad \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}\Big|_{\Delta_s=0, V_s=\tau},$$

we obtain

$$\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}\Big|_{\Delta_s=0, V_s=V_s} \quad = \quad \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}, \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}\Big|_{\Delta_s=0, V_s=V_s},$$

and therefore

$$\int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O},\boldsymbol{\theta}^*,\Lambda_s^*)d\boldsymbol{b}\Big|_{\Delta_s=0} \;=\; \int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O},\boldsymbol{\theta}_0,\Lambda_{s0})d\boldsymbol{b}\Big|_{\Delta_s=0};$$

that is, (3.18) also holds for any V_s and $\Delta_s = 0$.

Thus, first to show that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$ and $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{b0}$, we let $\Delta_s = 0$ and $V_s = 0$ in (3.18). After integrating over $\boldsymbol{b}$, we have that, with probability one,

$$\begin{aligned}
&\int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O},\boldsymbol{\theta}^*,\Lambda_s^*)d\boldsymbol{b}\Big|_{\Delta_s=0,V_s=0} \;=\; \int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O},\boldsymbol{\theta}_0,\Lambda_{s0})d\boldsymbol{b}\Big|_{\Delta_s=0,V_s=0}\\
\Rightarrow\quad &\int_{\boldsymbol{b}} \exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}^* + \widetilde{\boldsymbol{X}}_j\boldsymbol{b}) - B(\boldsymbol{\beta}^*,\boldsymbol{b})}{A(D(t_j;\phi^*))} + C(Y_j;D(t_j;\phi^*))\right]\right\}\\
&\qquad\qquad \times (2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b^*|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_b^{*-1}\boldsymbol{b}\right\}d\boldsymbol{b}\\
&= \int_{\boldsymbol{b}} \exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_j\boldsymbol{b}) - B(\boldsymbol{\beta}_0,\boldsymbol{b})}{A(D(t_j;\phi_0))} + C(Y_j;D(t_j;\phi_0))\right]\right\}\\
&\qquad\qquad \times (2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_{b0}|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}\right\}d\boldsymbol{b}\\
\Rightarrow\quad &\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))} + C(Y_j;D(t_j;\phi^*))\right]\right\}\times|\boldsymbol{\Sigma}_b^*|^{-1/2}\\
&\qquad\qquad \times\int_{\boldsymbol{b}}\exp\left\{\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j\boldsymbol{b}}{A(D(t_j;\phi^*))} - \sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b})}{A(D(t_j;\phi^*))} - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_b^{*-1}\boldsymbol{b}\right\}d\boldsymbol{b}\\
&= \exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}_0}{A(D(t_j;\phi_0))} + C(Y_j;D(t_j;\phi_0))\right]\right\}\times|\boldsymbol{\Sigma}_{b0}|^{-1/2}\\
&\qquad\qquad \times\int_{\boldsymbol{b}}\exp\left\{\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j\boldsymbol{b}}{A(D(t_j;\phi_0))} - \sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))} - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}\right\}d\boldsymbol{b}.
\end{aligned}$$

The left hand side becomes

$$\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))} + C(Y_j;D(t_j;\phi^*))\right]\right\}\times|\boldsymbol{\Sigma}_b^*|^{-1/2}$$

$$
\begin{aligned}
&\times \int_{\boldsymbol{b}} \exp\Bigg\{ -\frac{1}{2}\Bigg[\big(\Sigma_b^{*-1/2}\boldsymbol{b}\big)^T\big(\Sigma_b^{*-1/2}\boldsymbol{b}\big) - 2\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\boldsymbol{b} \\
&\qquad\qquad + \Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)^T\Bigg] \\
&\qquad + \frac{1}{2}\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)^T - \sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b})}{A(D(t_j;\phi^*))}\Bigg\} d\boldsymbol{b} \\
&= \exp\Bigg\{\sum_{j=1}^{n_N}\Bigg[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))} + C(Y_j;D(t_j;\phi^*))\Bigg] \\
&\qquad + \frac{1}{2}\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)^T\Bigg\} \times |\boldsymbol{\Sigma}_b^*|^{-1/2} \\
&\quad \times \int_{\boldsymbol{b}} \exp\Bigg\{ -\frac{1}{2}\Bigg[\Sigma_b^{*-1/2}\boldsymbol{b} - \Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)^T\Bigg]^T\Bigg[\Sigma_b^{*-1/2}\boldsymbol{b} - \Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)^T\Bigg]\Bigg\} \\
&\qquad \times \exp\Bigg\{ -\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b})}{A(D(t_j;\phi^*))}\Bigg\} d\boldsymbol{b} \\
&= \exp\Bigg\{\sum_{j=1}^{n_N}\Bigg[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))} + C(Y_j;D(t_j;\phi^*))\Bigg] \\
&\qquad + \frac{1}{2}\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)^T\Bigg\} \times \mathrm{E}\Bigg[\exp\Bigg\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b})}{A(D(t_j;\phi^*))}\Bigg\}\Bigg].
\end{aligned}
\tag{3.19}
$$

Likewise, the right-hand side becomes

$$
\begin{aligned}
&\exp\Bigg\{\sum_{j=1}^{n_N}\Bigg[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}_0}{A(D(t_j;\phi_0))} + C(Y_j;D(t_j;\phi_0))\Bigg] + \frac{1}{2}\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\Bigg)\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\Bigg)^T\Bigg\} \\
&\quad \times \mathrm{E}\Bigg[\exp\Bigg\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\Bigg\}\Bigg].
\end{aligned}
\tag{3.20}
$$

Then, to compare the coefficients of $\boldsymbol{Y}^T\boldsymbol{Y}$ and $\boldsymbol{Y}$ in the exponential part and the constant term out of the exponential part from (3.19) and (3.20), we have

$$
\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\Bigg)^T = \Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\Bigg)\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\Bigg)^T,
\tag{3.21}
$$

$$\sum_{j=1}^{n_N} \frac{Y_j \boldsymbol{X}_j \boldsymbol{\beta}^*}{A(D(t_j;\phi^*))} = \sum_{j=1}^{n_N} \frac{Y_j \boldsymbol{X}_j \boldsymbol{\beta}_0}{A(D(t_j;\phi_0))}, \tag{3.22}$$

and

$$\mathrm{E}\left[\exp\left\{ -\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}^*;\boldsymbol{b})}{A(D(t_j;\phi^*))} \right\} \right] = \mathrm{E}\left[\exp\left\{ -\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))} \right\} \right] \tag{3.23}$$

Define $\widetilde{\boldsymbol{X}}_j^* = \widetilde{\boldsymbol{X}}_j / A(D(t_j;\phi^*))$ and $\widetilde{\boldsymbol{X}}_{j0} = \widetilde{\boldsymbol{X}}_j / A(D(t_j;\phi_0))$ and $\widetilde{\boldsymbol{X}}^* = (\widetilde{\boldsymbol{X}}_1^{*T}, \ldots, \widetilde{\boldsymbol{X}}_{n_N}^{*T})^T$ and $\widetilde{\boldsymbol{X}}_0 = (\widetilde{\boldsymbol{X}}_{10}^{*T}, \ldots, \widetilde{\boldsymbol{X}}_{n_N 0}^{T})^T$. Then, (3.21) can be expressed as

$$\boldsymbol{Y}^T \widetilde{\boldsymbol{X}}^* \widetilde{\boldsymbol{X}}^{*T} \boldsymbol{Y} \;=\; \boldsymbol{Y}^T \widetilde{\boldsymbol{X}}_0 \widetilde{\boldsymbol{X}}_0^T \boldsymbol{Y},$$

and we obtain $\widetilde{\boldsymbol{X}}^* \widetilde{\boldsymbol{X}}^{*T} = \widetilde{\boldsymbol{X}}_0 \widetilde{\boldsymbol{X}}_0^T$ for the coefficients of $\boldsymbol{Y}^T \boldsymbol{Y}$. For the j-th diagonal element, we have

$$\widetilde{\boldsymbol{X}}_j^* \widetilde{\boldsymbol{X}}_j^{*T} = \widetilde{\boldsymbol{X}}_{j0} \widetilde{\boldsymbol{X}}_{j0}^T \quad \Rightarrow \quad \frac{\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))} \frac{\widetilde{\boldsymbol{X}}_j^{*T}}{A(D(t_j;\phi^*))} = \frac{\widetilde{\boldsymbol{X}}_{j0}}{A(D(t_j;\phi_0))} \frac{\widetilde{\boldsymbol{X}}_{j0}^T}{A(D(t_j;\phi_0))}.$$

By assumption (A5), $\big(A(D(t_j;\phi^*))\big)^2 = \big(A(D(t_j;\phi_0))\big)^2$. Then, we obtain $A(D(t_j;\phi^*)) = A(D(t_j;\phi_0))$ since both $A(D(t_j;\phi^*))$ and $A(D(t_j;\phi_0))$ are positive by the assumption for dispersion parameter of the generalized linear mixed model. By the continuous mapping theorem, we obtain $D(t_j;\phi^*) = D(t_j;\phi_0)$. By the similar argument, for the comparison of the coefficients of $\boldsymbol{Y}$, (3.22) can be written as

$$\boldsymbol{Y}^T \boldsymbol{X}^* \boldsymbol{\beta}^* = \boldsymbol{Y}^T \boldsymbol{X}_0 \boldsymbol{\beta}_0 \quad \Rightarrow \quad \boldsymbol{X}^* \boldsymbol{\beta}^* = \boldsymbol{X}_0 \boldsymbol{\beta}_0,$$

where the j-the elements $(\boldsymbol{X}_j / A(D(t_j;\phi^*)))\boldsymbol{\beta}^* = (\boldsymbol{X}_j / A(D(t_j;\phi_0)))\boldsymbol{\beta}_0$. By the result $A(D(t_j;\phi^*)) = A(D(t_j;\phi_0))$ and assumption (A5), we obtain $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. In (3.23) for the constant term, note that the random effect $\boldsymbol{b}$ on the left-hand side follows a multivariate normal distribution with mean $\boldsymbol{\Sigma}_b^{*1/2}\big(\sum_{j=1}^{n_N} Y_j \widetilde{\boldsymbol{X}}_j / A(D(t_j;\phi^*))\big)^T$ and covariance $\boldsymbol{\Sigma}_b^*$ while the random effect $\boldsymbol{b}$ on the right-hand side follows a multivariate normal distri-

bution with mean $\boldsymbol{\Sigma}_{b0}^{1/2}\big(\sum_{j=1}^{n_N} Y_j \widetilde{\boldsymbol{X}}_j / A(D(t_j;\phi_0))\big)^T$ and covariance $\boldsymbol{\Sigma}_{b0}$. Since $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$, $\sum_{j=1}^{n_N} B(\boldsymbol{\beta}^*;\boldsymbol{b})/A(D(t_j;\phi^*)) = \sum_{j=1}^{n_N} B(\boldsymbol{\beta}_0;\boldsymbol{b})/A(D(t_j;\phi_0))$. Thus, to hold the equality of the expected values in (3.23), the random effects $\boldsymbol{b}$ on both sides follow the same distribution; that is, $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{b0}$.

Next, to show that $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ and $\Lambda_s^* = \Lambda_{s0}$, we let $\Delta_s = 0$ in (3.18). Through the similar arguments done for the proof of $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$ and $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{b0}$, we obtain

$$\begin{aligned}
&\mathrm{E}\bigg[\exp\bigg\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b})}{A(D(t_j;\phi^*))} - \int_0^{V_s}\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}^*\circ\boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}^*\big\}d\Lambda_s^*(t)\bigg\}\bigg] \\
&= \mathrm{E}\bigg[\exp\bigg\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))} - \int_0^{V_s}\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\big\}d\Lambda_{s0}(t)\bigg\}\bigg], \quad (3.24)
\end{aligned}$$

where the random effects $\boldsymbol{b}$ follow a multivariate normal distribution with mean $\boldsymbol{\Sigma}_{b0}^{1/2}\big(\sum_{j=1}^{n_N} Y_j \widetilde{\boldsymbol{X}}_j / A(D(t_j;\phi_0))\big)^T$ and covariance $\boldsymbol{\Sigma}_{b0}$. For any fixed $\widetilde{\boldsymbol{X}}$, treating $\widetilde{\boldsymbol{X}}^T\boldsymbol{Y}$ as a parameter in this normal family, $\boldsymbol{b}$ is the complete statistic for $\widetilde{\boldsymbol{X}}^T\boldsymbol{Y}$. Therefore,

$$\begin{aligned}
&\exp\bigg\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b})}{A(D(t_j;\phi^*))} - \int_0^{V_s}\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}^*\circ\boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}^*\big\}d\Lambda_s^*(t)\bigg\} \\
&= \exp\bigg\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))} - \int_0^{V_s}\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\big\}d\Lambda_{s0}(t)\bigg\}.
\end{aligned}$$

Since $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$, equivalently, we have

$$\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}^*\circ\boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}^*\big\}\lambda_s^*(t) = \exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\big\}\lambda_{s0}(t).$$

By assumptions (A2) and (A5), $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ and $\Lambda_s^* = \Lambda_{s0}$.

Since all the three steps are completed, we can conclude that, with probability one, $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\Lambda}}$ converges to $\boldsymbol{\Lambda}_0$ in $[0,\tau]$. Moreover, as mentioned in the beginning of this proof for consistency, since $\boldsymbol{\Lambda}_0$ is continuous in $[0,\tau]$, the latter can be strengthened to uniform convergence; that is, $\sup_{t\in[0,\tau]}\|\widehat{\boldsymbol{\Lambda}}(t) - \boldsymbol{\Lambda}_0(t)\| \to 0$ almost

surely. Therefore, Theorem 3.1 is proved.

3.5.2 Proof of asymptotic normality

Asymptotic distribution for the proposed estimator can be shown if we can verify the conditions of Theorem 3.3.1 (p310) in van der Vaart and Wellner (1996). Then, we will show that the distribution is normal. For completeness, we state this theorem below following Theorem 4 in Appendix A of Parner (1998).

Theorem 3.3. (Theorem 3.3.1 in van der Vaart and Wellner, 1996) *Let U_n and U be random maps and a fixed map, respectively, from ξ to a Banach space such that:*

(a) $\sqrt{n}(U_n - U)(\widehat{\xi}_n) - \sqrt{n}(U_n - U)(\xi_0) = o_P^*(1 + \sqrt{n}\|\widehat{\xi}_n - \xi_0\|)$.

(b) The sequence $\sqrt{n}(U_n - U)(\xi_0)$ *converges in distribution to a tight random element* $\boldsymbol{W}$.

(c) the function $\xi \to U(\xi)$ *is Fréchet differentiable at* ξ_0 *with a continuously invertible derivative* ∇U_{ξ_0} *(on its range).*

(d) U_{ξ_0} *and* $\widehat{\xi}_n$ *satisfies* $U_n(\widehat{\xi}_n) = o_P^*(n^{-1/2})$ *and converges in outer probability to* ξ_0.

Then $\sqrt{n}(\widehat{\xi}_n - \xi_0) \Rightarrow \nabla U_{\xi_0}^{-1}\boldsymbol{W}$.

We will prove the conditions (a)~(d). In our situation, the parameter $\xi_s = (\boldsymbol{\theta}, \boldsymbol{\Lambda}_s) \in \Xi = \{(\boldsymbol{\theta}, \boldsymbol{\Lambda}_s) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda_s(t) - \Lambda_{s0}(t)| \le \delta,\ s = 1, \ldots, S\}$ for a fixed small constant δ. We note that Ξ is a convex set. Define a set $\mathscr{H} = \{(\mathbf{h}_1, h_2) : \|\mathbf{h}_1\| \le 1, \|h_2\|_V \le 1\}$, where $\|h_2\|_V$ is the total variation of h_2 in $[0, \tau]$ defined as

$$\sup_{0=t_0\le t_2\le\cdots\le t_k=\tau} \sum_{j=1}^{k} |h_2(t_j) - h_2(t_{j-1})|.$$

Furthermore, we define that, for stratum s,

$$\begin{aligned} U_{m_s}(\xi_s)(\mathbf{h}_1, h_2) &= \mathbf{P}_{m_s}\{l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2]\} \\ \text{and} \qquad U_s(\xi_s)(\mathbf{h}_1, h_2) &= \mathbf{P}\{l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2]\}, \end{aligned}$$

where $l_\theta(\boldsymbol{\theta}, \Lambda_s)$ is the first derivative of the log-likelihood function from one single subject belonging to stratum s, denoted by $l(\boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)$, with respect to $\boldsymbol{\theta}$, and $l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)$ is the derivative of $l(\boldsymbol{O}; \boldsymbol{\theta}, \Lambda_{s\varepsilon})$ at $\varepsilon = 0$, where $\Lambda_{s\varepsilon}(t) = \int_0^t (1 + \varepsilon h_2(u)) d\Lambda_{s0}(u)$. Therefore, we can see that both U_{m_s} and U_s map from Ξ to $\ell^\infty(\mathscr{H})$ and $\sqrt{m_s}\{U_{m_s}(\xi_s) - U_s(\xi_s)\}$ is an empirical process in the space $\ell^\infty(\mathscr{H})$.

Denote $(\mathbf{h}_1^\beta, \mathbf{h}_1^\phi, \mathbf{h}_1^b, \mathbf{h}_1^\psi, \mathbf{h}_1^\gamma)$ as the corresponding components of $\mathbf{h}_1$ for the parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathrm{Vec}(\boldsymbol{\Sigma}_b), \boldsymbol{\psi}, \boldsymbol{\gamma})$, respectively. From Section 3.5.3.2, for any $(\mathbf{h}_1, h_2) \in \mathscr{H}$, the class

$$\begin{aligned} \mathscr{G} = \Big\{ & l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2], \\ & \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda_s(t) - \Lambda_{s0}(t)| \le \delta,\ (\mathbf{h}_1, h_2) \in \mathscr{H} \Big\} \end{aligned}$$

is shown as P-Donsker (Section 2.1 of van der Vaart and Wellner (1996), and it is also implied that

$$\sup_{(\mathbf{h}_1, h_2)\in\mathscr{H}} \mathbf{P}\left[l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2] \right]^2 \longrightarrow 0$$

as $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda_s(t) - \Lambda_{s0}(t)| \to 0$. Then, we conclude the followings:

(a) follows from Lemma 3.3.5 (p311) of van der Vaart and Wellner (1996).

(b) holds as a result of Section 3.5.3.2 and the convergence is defined in the metric space $\ell^\infty(\mathscr{H})$ by the Donsker theorem (Section 2.5 of van der Vaart and Wellner

(1996).

(d) is true because $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)$ maximizes $\mathbf{P}_{m_s} l(\boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)$, $(\boldsymbol{\theta}_0, \Lambda_{s0})$ maximizes $\mathbf{P}\, l(\boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)$, and $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)$ converges to $(\boldsymbol{\theta}_0, \Lambda_{s0})$ from Theorem 3.1.

Now, we need to verify the conditions in (c). Since the proof of the first half in (c), that the function $\xi \to U(\xi)$ is *Fréchet* differentiable at ξ_0, is given in Section 3.5.3.3, we will only prove that the derivative ∇U_{ξ_0} is continuously invertible on its range $\ell^\infty(\mathscr{H})$. According to Section 3.5.3.3, ∇U_{ξ_0} can be expressed as follows: for any $(\boldsymbol{\theta}_1, \Lambda_{s1})$ and $(\boldsymbol{\theta}_1, \Lambda_{s1})$ in Ξ,

$$\nabla U_{\xi_0}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \Lambda_{s1} - \Lambda_{s2})[\mathbf{h}_1, h_2] = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T \Omega_1[\mathbf{h}_1, h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1, h_2] d(\Lambda_{s1} - \Lambda_{s2})(t), \tag{3.25}$$

where both Ω_1 and Ω_2 are linear operators on $\mathscr{H}$, and $\Omega = (\Omega_1, \Omega_2)$ maps $\mathscr{H} \subset \boldsymbol{R}^d \times \mathrm{BV}[0, \tau]$ to $\boldsymbol{R}^d \times \mathrm{BV}[0, \tau]$, where $\mathrm{BV}[0, \tau]$ contains all the functions with finite total variation in $[0, \tau]$. The explicit expressions of Ω_1 and Ω_2 are given in Section 3.5.3.3. From (3.25), we can treat $(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \Lambda_{s1} - \Lambda_{s2})$ as an element in $\ell^\infty(\mathscr{H})$ via the following definition:

$$(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \Lambda_{s1} - \Lambda_{s2})[\mathbf{h}_1, h_2] = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^T \mathbf{h}_1 + \int_0^\tau h_2(t) d(\Lambda_{s1} - \Lambda_{s2})(t), \quad \forall (\mathbf{h}_1, h_2) \in \boldsymbol{R}^d \times \mathrm{BV}[0, \tau].$$

Then ∇U_{ξ_0} can be expanded as a linear operator from $\ell^\infty(\mathscr{H})$ to itself. Therefore, if we can show that there exists some positive constant ε such that $\varepsilon \mathscr{H} \subset \Omega(\mathscr{H})$, then we will have that for any $(\delta\boldsymbol{\theta}, \delta\Lambda_s) \in \ell^\infty(\mathscr{H})$,

$$\begin{aligned} \|\nabla U_{\xi_0}(\delta\boldsymbol{\theta}, \delta\Lambda_s)\|_{\ell^\infty(\mathscr{H})} &= \sup_{(\mathbf{h}_1, h_2) \in \mathscr{H}} \left| \delta\boldsymbol{\theta}^T \Omega_1[\mathbf{h}_1, h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1, h_2] d\delta\Lambda_s(t) \right| \\ &= \|(\delta\boldsymbol{\theta}, \delta\Lambda_s)\|_{\ell^\infty(\Omega(\mathscr{H}))} \geq \varepsilon \|(\delta\boldsymbol{\theta}, \delta\Lambda_s)\|_{\ell^\infty(\mathscr{H})}, \end{aligned}$$

and ∇U_{ξ_0} will be continuously invertible.

Note that to prove $\varepsilon\mathscr{H} \subset \Omega(\mathscr{H})$ for some ε is equivalent to showing that Ω is invertible. We also note from Section 3.5.3.3, that Ω is the summation of an invertible operator and a compact operator. By Theorem 4.25 of Rudin (1973), for the proof of the invertibility of Ω, it is sufficient to verify that Ω is one to one: if $\Omega[\mathbf{h}_1, h_2] = 0$, then, by choosing $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 = \varepsilon^*\mathbf{h}_1$ and $\Lambda_{s1} - \Lambda_{s2} = \varepsilon^* \int h_2 d\Lambda_{s0}$ in (3.25) for a small constant ε^*, we obtain

$$\nabla U_{\xi_0}(\mathbf{h}_1, \int h_2 d\Lambda_{s0})[\mathbf{h}_1, h_2] = \varepsilon^*(\mathbf{h}_1^T, h_2)\begin{pmatrix} \Omega_1[\mathbf{h}_1, h_2] \\ \Omega_2[\mathbf{h}_1, h_2] \end{pmatrix} = \varepsilon^*(\mathbf{h}_1^T, h_2)\Omega[\mathbf{h}_1, h_2] = 0.$$

By the definition of ∇U_{ξ_0}, we note that $\nabla U_{\xi_0}(\mathbf{h}_1, \int h_2 d\Lambda_{s0})[\mathbf{h}_1, h_2]$ is the negative information matrix in the submodel $(\boldsymbol{\theta}_0 + \varepsilon\mathbf{h}_1, \Lambda_{s0} + \varepsilon \int h_2 d\Lambda_{s0})$. Thus, the score function along this submodel should be zero with probability one; that is, $l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2] = 0$; that is, with probability one, for the numerator of the score function

$$\begin{aligned}
0 = &\int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \times \left[\frac{\boldsymbol{b}^T \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{b}}{2} - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b\right) \right. \\
&+ \sum_{j=1}^{n_N} \left\{ - \left(\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_j\boldsymbol{b}) - B(\boldsymbol{\beta}_0; \boldsymbol{b})}{A(D(t_j; \phi_0))^2} \right) \left(A'(D(t_j; \phi_0)) \right) \mathrm{h}_1^\phi + C'(Y_j; D(t_j; \phi_0))\ \mathrm{h}_1^\phi \right\} \\
&+ \sum_{j=1}^{n_N} \left\{ \frac{Y_j \boldsymbol{X}_j}{A(D(t_j; \phi_0))} \mathbf{h}_1^\beta - B'(\boldsymbol{\beta}_0; \boldsymbol{b}) \mathbf{h}_1^\beta \right\} + \Delta_s \left\{ \left(\widetilde{\boldsymbol{Z}}(V_s) \circ \mathbf{h}_1^\psi \right)^T \boldsymbol{b} + \boldsymbol{Z}(V_s) \mathbf{h}_1^\gamma \right\} \\
&\left. - \int_0^{V_s} \exp\left\{ \widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0 \circ \boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0 \right\} \times \left\{ (\widetilde{\boldsymbol{Z}}(t) \circ \mathbf{h}_1^\psi)^T \boldsymbol{b} + \boldsymbol{Z}(t)\mathbf{h}_1^\gamma \right\} d\Lambda_{s0}(t) \right] d\boldsymbol{b} \\
&+ \int_{\boldsymbol{b}} G(\boldsymbol{b}, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \times \left[\Delta_s h_2(V_s) - \int_0^{V_s} h_2(t) \exp\left\{ \widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi} \circ \boldsymbol{b}) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0 \right\} d\Lambda_{s0}(t) \right] d\boldsymbol{b},
\end{aligned} \tag{3.26}$$

where $A'(D(t_j; \phi_0))$and $C'(Y_j; D(t_j; \phi_0))$ are the derivatives of $A(\phi_j)$and $C(Y_j; \phi_j)$ with

respect to ϕ_j evaluated at $D(t_j;\phi_0)$ and $B'(\boldsymbol{\beta}_0;\boldsymbol{b})$ is the derivative of $B(\boldsymbol{\beta};\boldsymbol{b})$ with respect to $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta}_0$. Note that (3.26) holds with probability one, so it may not hold for any $V_s \in [0,\tau]$ when $\Delta_s = 0$. However, by the similar arguments done in Section 3.5.1, if we integrate both sides from V_s to τ and substract the obtained equation from (3.26) at $\Delta_s = 0$ and $V_s = \tau$, it is easily shown that (3.26) also holds for any $V_s \in [0,\tau]$ when $\Delta_s = 0$. Hence, the proof of the invertibility of Ω will be completed if we can show $\mathbf{h}_1 = 0$ and $h_2(t) = 0$ from (3.26).

To show $\mathbf{h}_1 = 0$, particularly we let $\Delta_s = 0$ and $V_s = 0$ in (3.26) and obtain

$$
\begin{aligned}
0 = \int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) \times \Bigg[& \frac{\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}}{2} - \frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\right) \\
& + \sum_{j=1}^{n_N}\Bigg\{ -\left(\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}_0+\widetilde{\boldsymbol{X}}_j\boldsymbol{b}) - B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))^2}\right)\left(A'(D(t_j;\phi_0))\right)\,\mathrm{h}_1^{\phi} + C'(Y_j;D(t_j;\phi_0))\,\mathrm{h}_1^{\phi}\Bigg\} \\
& + \sum_{j=1}^{n_N}\left\{\frac{Y_j\boldsymbol{X}_j}{A(D(t_j;\phi_0))}\mathbf{h}_1^{\beta} - B'(\boldsymbol{\beta}_0;\boldsymbol{b})\mathbf{h}_1^{\beta}\right\}\Bigg] d\boldsymbol{b} \\
= \mathrm{E}\Bigg[& \exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\} \times \frac{\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}}{2}\Bigg] \\
& + \mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\right] \times \Bigg\{ -\frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b) \\
& \qquad + \sum_{j=1}^{n_N}\left(-\left(\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}_0}{(A(D(t_j;\phi_0)))^2}\right)(A'(D(t_j;\phi_0)))\,\mathrm{h}_1^{\phi} + C'(Y_j;D(t_j;\phi_0))\,\mathrm{h}_1^{\phi}\right) \\
& \qquad + \sum_{j=1}^{n_N}\left(-\left(\frac{Y_j\boldsymbol{X}_j}{A(D(t_j;\phi_0))}\right)\mathbf{h}_1^{\beta}\right)\Bigg\} \\
& + \mathrm{E}\Bigg[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\} \times \sum_{j=1}^{n_N}\Bigg\{ -\left(\frac{Y_j\widetilde{\boldsymbol{X}}_j\boldsymbol{b} - B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))^2}\right)\left(A'(D(t_j;\phi_0))\right)\,\mathrm{h}_1^{\phi} \\
& \qquad\qquad - B'(\boldsymbol{\beta}_0;\boldsymbol{b})\mathbf{h}_1^{\beta}\Bigg\}\Bigg]. \qquad (3.27)
\end{aligned}
$$

We first examine the coefficient for $\boldsymbol{Y}$ in (3.27).

$$
\begin{aligned}
&\mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\right] \\
&\qquad\times\sum_{j=1}^{n_N}\left\{-\frac{Y_j\boldsymbol{X}_j}{A(D(t_j;\phi_0))}\left(\frac{\boldsymbol{\beta}_0}{A(D(t_j;\phi_0))}A'(D(t_j;\phi_0))\ \mathrm{h}_1^{\phi}-\mathbf{h}_1^{\beta}\right)\right\} \\
&\quad+\sum_{j=1}^{n_N}\left\{-\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))^2}\big(A'(D(t_j;\phi_0))\big)\right\}\mathrm{h}_1^{\phi}\,\mathrm{E}\left[\boldsymbol{b}\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\right] \\
&=\sum_{j=1}^{n_N}\left\{-\frac{Y_j}{A(D(t_j;\phi_0))}\left[\boldsymbol{X}_j\left(\frac{\boldsymbol{\beta}_0}{A(D(t_j;\phi_0))}(A'(D(t_j;\phi_0)))\ \mathrm{h}_1^{\phi}-\mathbf{h}_1^{\beta}\right)\right.\right. \\
&\qquad\qquad\times\mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\right] \\
&\qquad\left.\left.+\frac{\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\big(A'(D(t_j;\phi_0))\big)\ \mathrm{h}_1^{\phi}\,\mathrm{E}\left[\boldsymbol{b}\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\right]\right]\right\} \\
&=\sum_{j=1}^{n_N}\left\{-\frac{Y_j}{A(D(t_j;\phi_0))}\left[\left[\boldsymbol{X}_j\left(\frac{\boldsymbol{\beta}_0}{A(D(t_j;\phi_0))}(A'(D(t_j;\phi_0)))\right)\mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\right]\right.\right.\right. \\
&\qquad\left.\left.+\frac{\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\big(A'(D(t_j;\phi_0))\big)\,\mathrm{E}\left[\boldsymbol{b}\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\right]\right]\mathrm{h}_1^{\phi} \\
&\qquad\left.-\left[\boldsymbol{X}_j\,\mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\right]\right]\mathbf{h}_1^{\beta}\right]\right\} \\
&=0
\end{aligned}
$$

Based on assumption (A5), $\mathrm{h}_1^{\phi}=0$ and $\mathbf{h}_1^{\beta}=0$.

Then, we examine the constant terms without $\boldsymbol{Y}$ in (3.27). Since $\mathrm{h}_1^{\phi}=0$ and $\mathbf{h}_1^{\beta}=0$, (3.27) becomes

$$
\begin{aligned}
&\mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\times\frac{\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}}{2}\right] \\
&\qquad+\mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\times\left\{-\frac{1}{2}\,\mathrm{Tr}(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b)\right\}\right] \\
&=\ \mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\times\left\{\frac{\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}}{2}-\frac{1}{2}\,\mathrm{Tr}(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b)\right\}\right]\ =\ 0,
\end{aligned}
$$

where $\boldsymbol{b}$ follows a multivariate normal distribution with mean $\boldsymbol{\Sigma}_{b0}^{1/2}\big[\sum_{j=1}^{n_N}\big(Y_j\widetilde{\boldsymbol{Z}}_j/A(D(t_j;$

$\phi_0)))\big]^T$ and covariance $\boldsymbol{\Sigma}_{b0}$. For any fixed $\widetilde{\boldsymbol{X}}$, treating $\boldsymbol{X}^T\boldsymbol{Y}$ as a parameter in this normal family, $\boldsymbol{b}$ is the complete statistic for $\boldsymbol{X}^T\boldsymbol{Y}$, therefore,

$$\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}\right\}\times\left\{\frac{\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}}{2}-\frac{1}{2}\operatorname{Tr}(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b)\right\} \;=\; 0.$$

Since $\exp\big\{-\sum_{j=1}^{n_N}\big(B(\boldsymbol{\beta}_0;\boldsymbol{b})/A(D(t_j;\phi_0))\big)\big\}\neq 0$, $\big[\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}-\operatorname{Tr}(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b)\big]/2=0$ by (A5). Then, since $\boldsymbol{\Sigma}_{b0}^{-1}\neq 0$, then $\boldsymbol{D}_b=0$.

Next, we let $\Delta_s=0$ in (3.26) and obtain

$$\begin{aligned}
0=&\int_{\boldsymbol{b}}G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\times\left[\frac{\boldsymbol{b}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}}{2}-\frac{1}{2}\operatorname{Tr}\big(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\big)\right.\\
&+\sum_{j=1}^{n_N}\left\{-\left(\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}_0+\widetilde{\boldsymbol{X}}_j\boldsymbol{b})-B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))^2}\right)\big(A'(D(t_j;\phi_0))\big)\ \mathrm{h}_1^{\phi}+C'(Y_j;D(t_j;\phi_0))\ \mathrm{h}_1^{\phi}\right\}\\
&+\sum_{j=1}^{n_N}\left\{\frac{Y_j\boldsymbol{X}_j}{A(D(t_j;\phi_0))}\mathbf{h}_1^{\beta}-B'(\boldsymbol{\beta}_0;\boldsymbol{b})\mathbf{h}_1^{\beta}\right\}\\
&-\left.\int_0^{V_s}\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\big\}\times\big\{(\widetilde{\boldsymbol{Z}}(t)\circ\mathbf{h}_1^{\psi})^T\boldsymbol{b}+\boldsymbol{Z}(t)\mathbf{h}_1^{\gamma}\big\}d\Lambda_{s0}(t)\right]d\boldsymbol{b}\\
&+\int_{\boldsymbol{b}}G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\times\left[-\int_0^{V_s}h_2(t)\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\big\}d\Lambda_{s0}(t)\right]d\boldsymbol{b}\\
=&\,\mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}-\int_0^{V_s}\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\big\}d\Lambda_{s0}(t)\right\}\right.\\
&\left.\times\int_0^{V_s}\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\big\}\times\big\{(\widetilde{\boldsymbol{Z}}(t)\circ\mathbf{h}_1^{\psi})^T\boldsymbol{b}+\boldsymbol{Z}(t)\mathbf{h}_1^{\gamma}+h_2(t)\big\}d\Lambda_{s0}(t)\right],
\end{aligned}\tag{3.28}$$

where $\boldsymbol{b}$ follows a multivariate normal distribution with mean $\boldsymbol{\Sigma}_{b0}^{1/2}\big[\sum_{j=1}^{n_N}\big(Y_j\widetilde{\boldsymbol{Z}}_j/A(D(t_j;\phi_0))\big)\big]^T$ and covariance $\boldsymbol{\Sigma}_{b0}$. Likewise, for any fixed $\widetilde{\boldsymbol{X}}$, treating $\boldsymbol{X}^T\boldsymbol{Y}$ as a parameter in this normal family, $\boldsymbol{b}$ is the complete statistic for $\boldsymbol{X}^T\boldsymbol{Y}$ and therefore

$$\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b})}{A(D(t_j;\phi_0))}-\int_0^{V_s}\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}d\Lambda_{s0}(t)\right\}$$
$$\times\int_0^{V_s}\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}\times\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\mathbf{h}_1^{\psi})^T\boldsymbol{b}+\boldsymbol{Z}(t)\mathbf{h}_1^{\gamma}+h_2(t)\right\}d\Lambda_{s0}(t)=0.$$

Since $\exp\left\{-\sum_{j=1}^{n_N}\left(B(\boldsymbol{\beta}_0;\boldsymbol{b})/A(D(t_j;\phi_0))\right)-\int_0^{V_s}\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}d\Lambda_{s0}(t)\right\}\neq 0$, equivalently

$$\int_0^{V_s}\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}\times\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\mathbf{h}_1^{\psi})^T\boldsymbol{b}+\boldsymbol{Z}(t)\mathbf{h}_1^{\gamma}+h_2(t)\right\}d\Lambda_{s0}(t)=0$$

by assumption (A5). From assumption (A5), this immediately gives $\mathbf{h}_1^{\psi}=0$, $\mathbf{h}_1^{\gamma}=0$ and $h_2(t)=0$. Hence, the proof of condition (c) is completed.

Since the conditions (a)–(d) have been proved, Theorem 3.3.1 of van der Vaart and Wellner (1996) concludes that $\sqrt{m_s}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}_s-\Lambda_{s0})$ weakly converges to a tight random element in $\ell^{\infty}(\mathscr{H})$. Furthermore, we obtain

$$\begin{aligned}&\sqrt{m_s}\nabla U_{\psi_0}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}_s-\Lambda_{s0})[\mathbf{h}_1,h_2]\\&\quad=\sqrt{m_s}(\mathbf{P}_{m_s}-\mathbf{P})\left\{l_{\theta}(\boldsymbol{\theta}_0,\Lambda_{s0})^T\mathbf{h}_1+l_{\Lambda_s}(\boldsymbol{\theta}_0,\Lambda_{s0})[h_2]\right\}+o_P(1),\end{aligned}\tag{3.29}$$

where $o_P(1)$ is a random variable which converges to zero in probability in $\ell^{\infty}(\mathscr{H})$. On the other hand, from (3.25), we have

$$\begin{aligned}&\sqrt{m_s}\nabla U_{\psi_0}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}_s-\Lambda_{s0})[\mathbf{h}_1,h_2]\\&\quad=\sqrt{m_s}\left\{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\Omega_1[\mathbf{h}_1,h_2]+\int_0^{\tau}\Omega_2[\mathbf{h}_1,h_2]d(\widehat{\Lambda}_s-\Lambda_{s0})(t)\right\}.\end{aligned}\tag{3.30}$$

By denoting $(\mathbf{h}_1^*,h_2^*)=\Omega^{-1}(\mathbf{h}_1,h_2)$, we have $(\mathbf{h}_1,h_2)=\Omega(\mathbf{h}_1^*,h_2^*)$, and replacing $(\mathbf{h}_1,h_2)$

with $(\mathbf{h}_1^*, h_2^*)$ in (3.29) and (3.30) leads to the followings, respectively.

$$\begin{aligned}
&\sqrt{m_s}\nabla U_{\psi_0}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda}_s - \Lambda_{s0})[\mathbf{h}_1^*, h_2^*] \\
&\quad = \sqrt{m_s}(\mathbf{P}_{m_s} - \mathbf{P})\big\{l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1^* + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2^*]\big\} + o_P(1),
\end{aligned}$$

and

$$\begin{aligned}
&\sqrt{m_s}\nabla U_{\psi_0}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda}_s - \Lambda_{s0})[\mathbf{h}_1^*, h_2^*] \\
&\quad = \sqrt{m_s}\left\{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\Omega_1[\mathbf{h}_1^*, h_2^*] + \int_0^\tau \Omega_2[\mathbf{h}_1^*, h_2^*]d(\widehat{\Lambda}_s - \Lambda_{s0})(t)\right\} \\
&\quad = \sqrt{m_s}\left\{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\mathbf{h}_1 + \int_0^\tau h_2(t)d(\widehat{\Lambda}_s - \Lambda_{s0})(t)\right\}.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
&\sqrt{m_s}\left\{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\mathbf{h}_1 + \int_0^\tau h_2(t)d(\widehat{\Lambda}_s - \Lambda_{s0})(t)\right\} \\
&\quad = \sqrt{m_s}(\mathbf{P}_{m_s} - \mathbf{P})\big\{l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1^* + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2^*]\big\} + o_P(1). \qquad (3.31)
\end{aligned}$$

Note that the first term on the right-hand side in (3.31) is $\sqrt{m_s}\{U_{m_s}(\boldsymbol{\theta}_0, \Lambda_{s0}) - U_s(\boldsymbol{\theta}_0, \Lambda_{s0})\}$, which is an empirical process in the space $\ell^\infty(\mathscr{H})$, and it is shown that $\mathscr{G}$ is P-Donsker in Section 3.5.3.2. Therefore, $\sqrt{m_s}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda}_s - \Lambda_{s0})$ weakly converges to a Gaussian process in $\ell^\infty(\mathscr{H})$.

In particular, if we choose $h_2 = 0$ in (3.31), then $\widehat{\boldsymbol{\theta}}^T\mathbf{h}_1$ is a asymptotic linear estimator for $\boldsymbol{\theta}_0^T\mathbf{h}_1$ with influence function being $l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1^* + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2^*]$. Since this influence function is in the linear space spanned by the score functions for $\boldsymbol{\theta}_0$ and Λ_{s0}, Proposition 3.3.1 (p65) in Bickel, Klaassen, Ritov and Wellner (1993) concludes that the influence function is the same as the efficient influence function for $\boldsymbol{\theta}_0^T\mathbf{h}_1$; that is $\widehat{\boldsymbol{\theta}}$ is an efficient estimator for $\boldsymbol{\theta}_0$ and Theorem 3.2 is proved.

3.5.3 Supplementary proofs

The proofs for P-Donsker property of the classes $\mathscr{F}$ and $\mathscr{G}$ needed in Sections 3.5.1 and 3.5.2 are presented in Sections 3.5.3.1~3.5.3.2 respectively. In Section 3.5.3.3, we prove *Fréchet* differentiability of $U(\xi)$ at ξ_0 and derive the derivative operator ∇U_{ξ_0} use in Section 3.5.2.

3.5.3.1 Proof of P-Donsker property of $\mathscr{F}$

We defined that a class $\mathscr{F} = \{Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) : v \in [0, \tau], \boldsymbol{\theta} \in \Theta, \Lambda_s \in \mathscr{A}, s = 1, \ldots, S\}$, where $\mathscr{A} = \{\Lambda_s \in \mathbb{W}, \Lambda_s(0) = 0, \Lambda_s(\tau) \leq B_{s0}, s = 1, \ldots, S\}$, B_{s0} is the constant given in the second step and $\mathbb{W}$ contains all nondecreasing functions in $[0, \tau]$. We can rewrite $Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)$ as

$$Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) = Q_1(v, \boldsymbol{O}; \boldsymbol{\theta}) \frac{Q_2(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)}{Q_3(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)},$$

where

$$\begin{aligned}
Q_1(v, \boldsymbol{O}; \boldsymbol{\theta}) &= \exp\left\{ \boldsymbol{Z}(v)\boldsymbol{\gamma} + \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j; \phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \right) (\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)^T \right. \\
&\qquad \left. + \frac{1}{2} (\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)^T \right\}, \\
Q_2(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) &= \int_{\boldsymbol{b}} \exp\left\{ -\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b} - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta}; \boldsymbol{b})}{A(D(t_j; \phi))} - \int_0^{V_s} \exp\left\{ (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b} + \boldsymbol{Z}(t)\boldsymbol{\gamma} \right.\right. \\
&\qquad \left.\left. + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2} \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j; \phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \right)^T + R(t) \right\} d\Lambda_s(t) \right\} d\boldsymbol{b}, \\
Q_3(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) &= \int_{\boldsymbol{b}} \exp\left\{ -\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b} - \sum_{j=1}^{n_N} \frac{B_2(\boldsymbol{\beta}; \boldsymbol{b})}{A(D(t_j; \phi))} - \int_0^{V_s} \exp\left\{ (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b} + \boldsymbol{Z}(t)\boldsymbol{\gamma} \right.\right. \\
&\qquad \left.\left. + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2} \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j; \phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \right)^T \right\} d\Lambda_s(t) \right\} d\boldsymbol{b},
\end{aligned}$$

$R(t) = (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)^T$, $B_1(\boldsymbol{\beta}; \boldsymbol{b}) = B(\boldsymbol{\beta}; g_1(\boldsymbol{b}))$, $B_2(\boldsymbol{\beta}; \boldsymbol{b}) = B(\boldsymbol{\beta}; g_2(\boldsymbol{b}))$, $g_1(\boldsymbol{b}) = \boldsymbol{\Sigma}_b^{1/2}\big[\boldsymbol{b} + \big(\sum_{j=1}^{n_N} \left(Y_j \widetilde{\boldsymbol{X}}_j / A(D(t_j; \phi)) \right) + (\Delta + 1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \big)^T \big]$ and $g_2(\boldsymbol{b}) = \boldsymbol{\Sigma}_b^{1/2}\big[\boldsymbol{b} +$

$\big(\sum_{j=1}^{n_N}\big(Y_j\widetilde{\boldsymbol{X}}_j/A(D(t_j;\phi))\big)+\Delta(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\big)^T\big]$.

Using assumption (A2), we can easily show that $Q_1(v,\boldsymbol{O};\boldsymbol{\theta})$ is continuously differentiable with respect to v and $\boldsymbol{\theta}$, and

$$\|\nabla_{\boldsymbol{\theta}}Q_1(v,\boldsymbol{O};\boldsymbol{\theta})\|+\left|\frac{d}{dv}Q_1(v,\boldsymbol{O};\boldsymbol{\theta})\right| \;\le\; e^{k_1+k_2\|\boldsymbol{Y}\|}$$

for some positive constants k_1 and k_2. Furthermore, it holds that

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}}Q_2(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\|+\left|\frac{d}{dv}Q_2(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\right| \\
&\quad\le \int_{\boldsymbol{b}}\left[\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}\right\}\times e^{k_3\|\boldsymbol{b}\|+k_4\|\boldsymbol{Y}\|+k_5}\times B_{s0}\right]d\boldsymbol{b} \\
&\quad\le e^{k_6+k_7\|\boldsymbol{Y}\|}
\end{aligned}$$

and $$\|\nabla_{\boldsymbol{\theta}}Q_3(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\|+\left|\frac{d}{dv}Q_3(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\right| \;\le\; e^{k_8+k_9\|\boldsymbol{Y}\|}$$

for some positive constants $k_3,k_4,\ldots,k_9$. Additionally, note that, for any $0<\Lambda<\infty$, $0<e^{-\Lambda}<1$ and $e^{-\Lambda}<\Lambda$ and thus $e^{-\Lambda_1}-e^{-\Lambda_2}<\Lambda_1-\Lambda_2$ for any Λ_1 and Λ_2 over $(0,\infty)$. Hence,

$$\begin{aligned}
&|Q_2(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_{s1})-Q_2(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_{s2})| \\
&=\left|\int_{\boldsymbol{b}}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}\right\}\times\left[\exp\left\{-\int_0^{V_s}\exp\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}\right.\right.\right. \\
&\quad\left.\left.+\boldsymbol{Z}(t)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\right)^T+R(t)\right\}d\Lambda_{s1}(t)\right\} \\
&\quad-\exp\left\{-\int_0^{V_s}\exp\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}+\boldsymbol{Z}(t)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{W}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}\right.\right.\right. \\
&\qquad\left.\left.\left.+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\right)^T+R(t)\right\}d\Lambda_{s2}(t)\right\}\Bigg]d\boldsymbol{b}\Bigg| \\
&\le\left|\int_{\boldsymbol{b}}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}\right\}\times\int_0^{V_s}\exp\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}+\boldsymbol{Z}(t)\boldsymbol{\gamma}\right.\right.
\end{aligned}$$

$$
\begin{aligned}
&+ (\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+R(t)\bigg\}d(\Lambda_{s1}-\Lambda_{s2})(t)\bigg]d\boldsymbol{b}\bigg| \\
&= \bigg|\int_{\boldsymbol{b}}\exp\bigg\{-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}\bigg\}\times\bigg[\int_0^{V_s}\exp\bigg\{-\frac{1}{2}\bigg[\boldsymbol{b}^T\boldsymbol{b}-2(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b} \\
&+ (\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T-(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T\bigg] \\
&+ \boldsymbol{Z}(t)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+R(t)\bigg\}d(\Lambda_{s1}-\Lambda_{s2})(t)\bigg]d\boldsymbol{b}\bigg| \\
&= \bigg|\int_{\boldsymbol{b}}\exp\bigg\{-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}\bigg\}\times(2\pi)^{d_b/2} \\
&\times(2\pi)^{-d_b/2}\times\bigg[\int_0^{V_s}\exp\bigg\{-\frac{1}{2}\bigg[\boldsymbol{b}-\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T\bigg]^T\bigg[\boldsymbol{b}-\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T\bigg]\bigg\} \\
&\quad\times\exp\bigg\{\frac{1}{2}(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T+\boldsymbol{Z}(t)\boldsymbol{\gamma} \\
&\quad+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+R(t)\bigg\}d(\Lambda_{s1}-\Lambda_{s2})(t)\bigg]d\boldsymbol{b}\bigg| \\
&\le \bigg|\mathrm{E}_b\bigg[\exp\bigg\{-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}\bigg\}\bigg](2\pi)^{d_b/2}\int_0^{V_s}\exp\bigg\{\frac{1}{2}(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T \\
&+ \boldsymbol{Z}(t)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+R(t)\bigg\}d(\Lambda_{s1}-\Lambda_{s2})(t)\bigg| \\
&= K_0\bigg|\int_0^{V_s}\exp\bigg\{\frac{1}{2}(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T \\
&+ \boldsymbol{Z}(t)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+R(t)\bigg\}d(\Lambda_{s1}-\Lambda_{s2})(t)\bigg| \\
&= K_0\bigg|-\int_0^{V_s}(\Lambda_{s1}(t)-\Lambda_{s2}(t))\frac{d}{dt}\bigg[\exp\bigg\{\frac{1}{2}(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T \\
&\quad+\boldsymbol{Z}(t)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+R(t)\bigg\}\bigg]d(t) \\
&+ (\Lambda_{s1}(V_s)-\Lambda_{s2}(V_s))\exp\bigg\{\frac{1}{2}(\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T \\
&\quad+\boldsymbol{Z}(V_s)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{\psi}^T)\bigg)^T+R(V_s)\bigg\}\bigg| \\
&\le K_0\bigg[\int_0^{V_s}\big|\Lambda_{s1}(t)-\Lambda_{s2}(t)\big|\bigg|\frac{d}{dt}\bigg[\exp\bigg\{\frac{1}{2}(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T
\end{aligned}
$$

$$
\begin{aligned}
&+\boldsymbol{Z}(t)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\Bigg)^T+R(t)\Bigg\}\Bigg]\Bigg|dt\\
&+\big|\Lambda_{s1}(V_s)-\Lambda_{s2}(V_s)\big|\exp\Bigg\{\frac{1}{2}(\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big((\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T\\
&\qquad+\boldsymbol{Z}(V_s)\boldsymbol{\gamma}+(\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{\psi}^T)\Bigg)^T+R(V_s)\Bigg\}\Bigg]\\
&\le e^{k_{10}+k_{11}\|\boldsymbol{Y}\|}\big\{\big|\Lambda_{s1}(V_s)-\Lambda_{s2}(V_s)\big|+\int_0^\tau\big|\Lambda_{s1}(t)-\Lambda_{s2}(t)\big|dt\big\},
\end{aligned}
$$

where $K_0=\mathrm{E}_b\big[\exp\big\{-\sum_{j=1}^{n_N}\big(B_1(\boldsymbol{\beta};\boldsymbol{b})/A(D(t_j;\phi))\big)\big\}\big](2\pi)^{d_b/2}$, k_{10} and k_{11} are positive constants. Similarly,

$$
\begin{aligned}
&|Q_3(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_{s1})-Q_3(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_{s2})|\\
&\quad\le\ e^{k_{12}+k_{13}\|\boldsymbol{Y}\|}\Bigg\{\big|\Lambda_{s1}(V_s)-\Lambda_{s2}(V_s)\big|+\int_0^\tau\big|\Lambda_{s1}(t)-\Lambda_{s2}(t)\big|dt\Bigg\},
\end{aligned}
$$

where k_{12} and k_{13} are positive constants.

On the other hand, there exist positive constants $k_{14},\ldots,k_{24}$ such that

$$
\begin{aligned}
&\big|Q_1(v,\boldsymbol{O};\boldsymbol{\theta})\big|\\
&=\Bigg|\exp\Bigg\{\boldsymbol{Z}(v)\boldsymbol{\gamma}+\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\Bigg)(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)^T\\
&\qquad+\frac{1}{2}(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)^T\Bigg\}\Bigg|\\
&\le e^{k_{14}+k_{15}\|\boldsymbol{Y}\|},\\
&\big|Q_2(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\big|\\
&=\Bigg|\int_{\boldsymbol{b}}\exp\Bigg\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}-\int_0^{V_s}\exp\Bigg\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}+\boldsymbol{Z}(t)\boldsymbol{\gamma}\\
&\qquad+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\Bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\Bigg)^T+R(t)\Bigg\}d\Lambda_s(t)\Bigg\}d\boldsymbol{b}\Bigg|\\
&\le\Bigg|\int_{\boldsymbol{b}}\exp\Bigg\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}\Bigg\}\times\Bigg[2\int_0^{V_s}\exp\Bigg\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}+\boldsymbol{Z}(t)\boldsymbol{\gamma}
\end{aligned}
$$

$$
\begin{aligned}
&+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\right)^T+R(t)\Big\}d\Lambda_s(t)\Big]d\boldsymbol{b}\Big| \\
&\le\left|\int_{\boldsymbol{b}}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}\right\}\times 2\ \exp\left\{k_{16}\|\boldsymbol{b}\|+k_{17}\|\boldsymbol{Y}\|+k_{18}\right\}\times B_{s0}\ d\boldsymbol{b}\right| \\
&\le e^{k_{19}+k_{20}\|\boldsymbol{Y}\|},
\end{aligned}
$$

and $Q_3(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)$

$$
\begin{aligned}
&=\int_{\boldsymbol{b}}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}-\sum_{j=1}^{n_N}\frac{B_2(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}-\int_0^{V_s}\exp\Big\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}+\boldsymbol{Z}(t)\boldsymbol{\gamma}\right. \\
&\qquad\left.+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\right)^T\Big\}d\Lambda_s(t)\Big\}d\boldsymbol{b}, \\
&\ge\int_{\boldsymbol{b}}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{b}-\sum_{j=1}^{n_N}\frac{B_2(\boldsymbol{\beta};\boldsymbol{b})}{A(D(t_j;\phi))}-\exp\left\{k_{21}\|\boldsymbol{b}\|+k_{22}\|\boldsymbol{Y}\|+k_{23}\right\}\times B_{s0}\right\}d\boldsymbol{b}, \\
&\ge k_{24}>0.
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}}Q(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\|+\left|\frac{d}{dv}Q(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\right| \\
&=\left\|(\nabla_{\boldsymbol{\theta}}Q_1)\frac{Q_2}{Q_3}+Q_1\left(\nabla_{\boldsymbol{\theta}}\frac{Q_2}{Q_3}\right)\right\|+\left|\left(\frac{d}{dv}Q_1\right)\frac{Q_2}{Q_3}+Q_1\left(\frac{d}{dv}\left(\frac{Q_2}{Q_3}\right)\right)\right| \\
&=\left\|(\nabla_{\boldsymbol{\theta}}Q_1)\frac{Q_2}{Q_3}+Q_1\left[(\nabla_{\boldsymbol{\theta}}Q_2)\frac{1}{Q_3}+Q_2\frac{(-1)}{Q_3^2}(\nabla_{\boldsymbol{\theta}}Q_3)\right]\right\| \\
&\qquad+\left|\left(\frac{d}{dv}Q_1\right)\frac{Q_2}{Q_3}+Q_1\left[\left(\frac{d}{dv}Q_2\right)\frac{1}{Q_3}+Q_2\frac{(-1)}{Q_3^2}\left(\frac{d}{dv}Q_3\right)\right]\right| \\
&=\left\|(\nabla_{\boldsymbol{\theta}}Q_1)\frac{Q_2}{Q_3}+(\nabla_{\boldsymbol{\theta}}Q_2)\frac{Q_1}{Q_3}-(\nabla_{\boldsymbol{\theta}}Q_3)\frac{Q_2}{Q_3^2}\right\|+\left|\left(\frac{d}{dv}Q_1\right)\frac{Q_2}{Q_3}+\left(\frac{d}{dv}Q_2\right)\frac{Q_1}{Q_3}-\left(\frac{d}{dv}Q_3\right)\frac{Q_1Q_2}{Q_3^2}\right| \\
&\le\left(\|\nabla_{\boldsymbol{\theta}}Q_1\|+\left|\frac{d}{dv}Q_1\right|\right)\left|\frac{Q_2}{Q_3}\right|+\left(\|\nabla_{\boldsymbol{\theta}}Q_2\|+\left|\frac{d}{dv}Q_2\right|\right)\left|\frac{Q_1}{Q_3}\right|+\left(\|\nabla_{\boldsymbol{\theta}}Q_3\|+\left|\frac{d}{dv}Q_3\right|\right)\left|\frac{Q_1Q_2}{Q_3^2}\right| \\
&\le e^{k_{25}+k_{26}\|\boldsymbol{Y}\|},
\end{aligned}
$$

for some positive constants k_{25} and k_{26}. Therefore, by the mean-value theorem, we conclude that, for any $(v_1,\boldsymbol{\theta}_1,\Lambda_{s1})$ and $(v_2,\boldsymbol{\theta}_2,\Lambda_{s2})$ in $[0,\tau]\times\Theta\times\mathscr{A}$,

$$\begin{aligned}
&\left|Q(v_1, \boldsymbol{O}; \boldsymbol{\theta}_1, \Lambda_{s1}) - Q(v_2, \boldsymbol{O}; \boldsymbol{\theta}_2, \Lambda_{s2})\right| \\
&\leq e^{k_{25}+k_{26}\|\boldsymbol{Y}\|}\left\{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \left|\Lambda_{s1}(V_s) - \Lambda_{s2}(V_s)\right| + \int_0^{V_s}\left|\Lambda_{s1}(t) - \Lambda_{s2}(t)\right|dt + |v_1 - v_2|\right\} \quad (3.32)
\end{aligned}$$

holds for some positive constants k_{25} and k_{26}.

Applying Theorem 2.7.5 (p159) in van der Vaart and Wellner (1996) to our situation, the entropy number for the class $\mathscr{A}$ satisfies $\log N_{[\cdot]}(\varepsilon, \mathscr{A}, L_2(P)) \leq K/\varepsilon$, where K is a constant. Thus, we can find $\exp\{K/\varepsilon\}$ brackets, $\{[L_j, U_j]\}$, to cover the class $\mathscr{A}$ such that $\|U_j - L_j\|_{L_2(P)} \leq \varepsilon$ for each pair of $[L_j, U_j]$. On the other hand, we can further find a partition of $[0, \tau] \times \Theta$, say $I_1 \bigcup I_2 \bigcup \cdots$, such that the number of partitions is of the order $(1/\varepsilon)^{d+1}$, and, for any $(v_1, \boldsymbol{\theta}_1)$ and $(v_2, \boldsymbol{\theta}_2)$ in the same partition, their Euclidean distance is less than ε. Therefore, the partition $\{I_1, I_2, \ldots\} \times \{[L_j, U_j]\}$ bracket covers $[0, \tau] \times \Theta \times \mathscr{A}$, and the total number of the partitions is of order $(1/\varepsilon)^{d+1}\exp\{1/\varepsilon\}$. Hence, from (3.32), for any I_l and $[L_j, U_j]$, the set of the functions $\{Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) : (v, \boldsymbol{\theta}) \in S_l, \Lambda_s \in \mathscr{A}, \Lambda_s \in [L_j, U_j]\}$ can be bracket covered by

$$\begin{aligned}
&\Bigg[Q(v_l, \boldsymbol{O}; \boldsymbol{\theta}_l, \Lambda_{sl}) - e^{k_{25}+k_{26}\|\boldsymbol{Y}\|}\left\{\varepsilon + \left|U_j(V_s) - L_j(V_s)\right| + \int_0^{V_s}\left|U_j(t) - L_j(t)\right|dt\right\}, \\
&\quad Q(v_l, \boldsymbol{O}; \boldsymbol{\theta}_l, \Lambda_{sl}) + e^{k_{25}+k_{26}\|\boldsymbol{Y}\|}\left\{\varepsilon + \left|U_j(V_s) - L_j(V_s)\right| + \int_0^{V_s}\left|U_j(t) - L_j(t)\right|dt\right\}\Bigg], \quad (3.33)
\end{aligned}$$

where $(v_l, \boldsymbol{\theta}_l)$ is a fixed point in I_l and Λ_{sj} is a fixed function in $[L_j, U_j]$. Note that the $L_2(P)$ distance between these two functions in the above bracket (3.33) is less than $O(\varepsilon)$. Therefore, we have

$$N_{[\cdot]}(\varepsilon, \mathscr{F}, \|\cdot\|_{L_2(P)} \leq O\left(\left(\frac{1}{\varepsilon}\right)^{d+1} e^{1/\varepsilon}\right).$$

Furthermore, $\mathscr{F}$ has an $L_2(P)$-integrable covering function, which is equal to $O($

$e^{k_{25}+k_{26}\|\boldsymbol{Y}\|}$). From Theorem 2.5.6 (p130) in van der Vaart and Wellner (1996), $\mathscr{F}$ is P-Donsker.

Additionally, in the above derivation, we also note that all the functions in $\mathscr{F}$ are bounded from below by $e^{-k_{27}-k_{28}\|\boldsymbol{Y}\|}$ for some positive constants k_{27} and k_{28}.

3.5.3.2 Proof of P-Donsker property of $\mathscr{G}$

Recall that we defined the class

$$\begin{aligned}\mathscr{G} \;=\; & \Big\{ l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0,\Lambda_{s0})^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}_0,\Lambda_{s0})[h_2], \\ & \qquad \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda_s(t)-\Lambda_{s0}(t)| \le \delta,\ (\mathbf{h}_1,h_2)\in\mathscr{H}\Big\},\end{aligned}$$

where $(\mathbf{h}_1^\beta,\mathbf{h}_1^\phi,\mathbf{h}_1^b,\mathbf{h}_1^\psi,\mathbf{h}_1^\gamma)$ denote the corresponding components of $\mathbf{h}_1$ for the parameters $(\boldsymbol{\beta},\boldsymbol{\phi},\mathrm{Vec}(\boldsymbol{\Sigma}_b),\boldsymbol{\psi},\boldsymbol{\gamma})$, respectively. We can write that for $(\mathbf{h}_1,h_2)\in\mathscr{H}$,

$$\begin{aligned}& l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2] \\ =& \left[\mu_1(\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 - \int_0^{V_s}\mu_2(t,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 d\Lambda_s(t)\right] + \Delta h_2(V_s) - \int_0^{V_s}\mu_3(t,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)h_2(t)d\Lambda_s(t),\end{aligned}$$

where

$$\begin{aligned}& \mu_1(\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 \\ =& \left\{\int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta},\Lambda_2)d\boldsymbol{b}\right\}^{-1} \times \int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta},\Lambda_2)\times\left[\frac{\boldsymbol{b}^T\boldsymbol{\Sigma}_b^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}}{2} - \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}_b^{-1}\boldsymbol{D}_b)\right. \\ & + \sum_{j=1}^{n_N}\left\{-\left(\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_j\boldsymbol{b}) - B(\boldsymbol{\beta};\boldsymbol{b})}{(A(D(t_j;\phi)))^2}\right)A'(D(t_j;\phi))h_1^{D(t_j;\phi)} + C'(Y_j;D(t_j;\phi))h_1^{D(t_j;\phi)}\right\} \\ & \left. + \sum_{j=1}^{n_N}\left(\frac{Y_j\boldsymbol{X}_j}{A(D(t_j;\phi))}\mathbf{h}_1^\beta - B'(\boldsymbol{\beta};\boldsymbol{b})\mathbf{h}_1^\beta\right) + \Delta_s\left[\left(\widetilde{\boldsymbol{Z}}(V_s)\circ\mathbf{h}_1^\psi\right)^T\boldsymbol{b} + \boldsymbol{Z}(V_s)\mathbf{h}_1^\gamma\right]\right]d\boldsymbol{b},\end{aligned}$$

$$
\begin{aligned}
&\mu_2(t,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1\\
&=\left\{\int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta},\Lambda_2)d\boldsymbol{b}\right\}^{-1}\\
&\quad\times\int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta},\Lambda_2)\times\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}\right\}\times\left[\left(\widetilde{\boldsymbol{Z}}(t)\circ\mathbf{h}_1^{\psi}\right)^T\boldsymbol{b}+\boldsymbol{Z}(t)\mathbf{h}_1^{\gamma}\right]d\boldsymbol{b},\\
&\mu_3(t,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\\
&=\left\{\int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta},\Lambda_2)d\boldsymbol{b}\right\}^{-1}\times\int_{\boldsymbol{b}} G(\boldsymbol{b},\boldsymbol{O};\boldsymbol{\theta},\Lambda_2)\times\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}\circ\boldsymbol{b})+\boldsymbol{Z}(t)\boldsymbol{\gamma}\right\}d\boldsymbol{b},
\end{aligned}
$$

$\boldsymbol{D}_b$ is the symmetric matrix such that $\mathrm{Vec}(\boldsymbol{D}_b)=\mathbf{h}_1^b$, $A'(D(t_j;\phi))$ and $C'(Y_j;D(t_j;\phi))$ are the derivatives of $A(D(t_j;\phi))$ and $C(Y_j;D(t_j;\phi))$ with respect to $D(t_j;\phi)$ respectively, and $B'(\boldsymbol{\beta};\boldsymbol{b})$ is the derivative of $B(\boldsymbol{\beta};\boldsymbol{b})$ with respect to $\boldsymbol{\beta}$.

For $l=1,2,3$, we denote $\nabla_{\boldsymbol{\theta}}\mu_l$ and $\nabla_{\Lambda_s}\mu_l[\delta\Lambda_s]$ as the derivatives of μ_l with respect to $\boldsymbol{\theta}$ and Λ_s along the path $\Lambda_s+\varepsilon\delta\Lambda_s$. Then, using the similar arguments done in Section 3.5.3.1, it is verified that $\nabla_{\Lambda_s}\mu_l[\delta\Lambda_s]=\int_0^t\mu_{l+3}(u,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)d\delta\Lambda_s(u)$ and there exist two positive constants q_1 and q_2 such that

$$
\sum_l\left\{|\mu_l|+|\nabla_{\boldsymbol{\theta}}\mu_l|\right\}\le e^{q_1+q_2\|\boldsymbol{Y}\|}
$$

By the mean value theorem, we have that, for any $(\boldsymbol{\theta},\Lambda_s,\mathbf{h}_1,h_2)$ and $(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s,\widetilde{\mathbf{h}}_1,\widetilde{h}_2)$ in $\Xi\times\mathscr{H}$,

$$
\begin{aligned}
&l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1+l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2]-l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T\widetilde{\mathbf{h}}_1-l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)[\widetilde{h}_2]\\
&=l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1+l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2]-l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T\mathbf{h}_1-l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)[h_2]\\
&\quad+l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T\mathbf{h}_1+l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)[h_2]-l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T\widetilde{\mathbf{h}}_1-l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)[\widetilde{h}_2]\\
&=\left[l_\theta(\boldsymbol{\theta},\Lambda_s)^T-l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T\right]\mathbf{h}_1+\left[l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)-l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)\right][h_2]\\
&\quad+l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T(\mathbf{h}_1-\widetilde{\mathbf{h}}_1)+l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)([h_2]-[\widetilde{h}_2])\\
&=(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^T\left[\frac{d}{d\boldsymbol{\theta}}l_{\boldsymbol{\theta}}(\boldsymbol{\theta},\Lambda_s)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\Lambda_s=\Lambda_s^*}\right]\mathbf{h}_1+\left[\frac{d}{d\Lambda_s}l_{\boldsymbol{\theta}}(\boldsymbol{\theta},\Lambda_s)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\Lambda_s=\Lambda_s^*}\right]^T[\Lambda_s-\widetilde{\Lambda}_s]\mathbf{h}_1
\end{aligned}
$$

$$
\begin{aligned}
&+ (\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^T\left[\frac{d}{d\boldsymbol{\theta}}l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\Lambda_s=\Lambda_s^*}\right][h_2] + \left[\frac{d}{d\Lambda_s}l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*,\Lambda_s=\Lambda_s^*}\right]^T[\Lambda_s-\widetilde{\Lambda}_s][h_2]\\
&+ l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T(\mathbf{h}_1-\widetilde{\mathbf{h}}_1) + l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)([h_2]-[\widetilde{h}_2])\\
=\ &(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^T\nabla_{\boldsymbol{\theta}}\mu_1(\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)\mathbf{h}_1 - (\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^T\int_0^{V_S}\nabla_{\boldsymbol{\theta}}\mu_2(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)d\Lambda_s^*(t)\mathbf{h}_1\\
&+\int_0^{V_s}\mu_4(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T\mathbf{h}_1 d(\Lambda_s-\widetilde{\Lambda}_s)(t)\\
&\quad-\int_0^{V_s}\int_0^t\mu_5(u,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T d(\Lambda_s-\widetilde{\Lambda}_s)(u)\mathbf{h}_1 d\Lambda_s^*(t)\\
&-\int_0^{V_s}\mu_2(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T(\Lambda_s-\widetilde{\Lambda}_s)\mathbf{h}_1 dt\\
&\quad-(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^T\int_0^{V_S}\nabla_{\boldsymbol{\theta}}\mu_3(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)h_2(t)d\Lambda_s^*(t)\\
&-\int_0^{V_s}\int_0^t\mu_6(u,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T d(\Lambda_s-\widetilde{\Lambda}_s)(u)h_2(t)d\Lambda_s^*(t)\\
&\quad-\int_0^{V_s}\mu_3(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T(\Lambda_s-\widetilde{\Lambda}_s)h_2(t)dt\\
&+\mu_1(\boldsymbol{O};\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T(\mathbf{h}_1-\widetilde{\mathbf{h}}_1) - \int_0^{V_S}\mu_2(t,\boldsymbol{O};\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T(\mathbf{h}_1-\widetilde{\mathbf{h}}_1)d\widetilde{\Lambda}_s(t)\\
&+\Delta_s(h_2(V_s)-\widetilde{h}_2(V_s)) - \int_0^{V_S}\mu_3(t,\boldsymbol{O};\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)(h_2(V_s)-\widetilde{h}_2(V_s))d\widetilde{\Lambda}_s(t), \qquad (3.34)
\end{aligned}
$$

where $(\boldsymbol{\theta}^*,\Lambda_s^*)$ is equal to $\varepsilon^*(\boldsymbol{\theta},\Lambda_s)+(1-\varepsilon^*)(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)$ for some $\varepsilon^*\in[0,1]$. Thus, we have

$$
\begin{aligned}
&\left|l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2] - l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T\widetilde{\mathbf{h}}_1 - l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)[\widetilde{h}_2]\right|\\
&\quad\le e^{q_1+q_2\|\boldsymbol{Y}\|}\Bigg\{\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\| + \|\mathbf{h}_1-\widetilde{\mathbf{h}}_1\| + |\Lambda_s(V_s)-\widetilde{\Lambda}_s(V_s)|\\
&\qquad\qquad+\int_0^\tau|\Lambda_s(t)-\widetilde{\Lambda}_s(t)|[dt+d|h_2(t)|+d|\widetilde{h}_2(t)|]\\
&\qquad\qquad+|h_2(V_s)-\widetilde{h}_2(V_s)| + \int_0^\tau|h_2(V_s)-\widetilde{h}_2(V_s)|[\Lambda_s(t)-\widetilde{\Lambda}_s(t)]\Bigg\}, \qquad (3.35)
\end{aligned}
$$

where $d|h_2(t)| = dh_2^+(t)+dh_2^-(t)$ and $d|\widetilde{h}_2(t)| = d\widetilde{h}_2^+(t)+d\widetilde{h}_2^-(t)$. As done in Section 3.5.3.1, by applying Theorem 2.7.5 (p159) in van der Vaart and Wellner (1996), we note that for a set $\mathscr{H}_2=\{h_2:\|h_2\|_V\le B_1\}$, $\log N_{[\cdot]}\big(\varepsilon,\mathscr{H}_2,L_2(P)\big)\le K/\varepsilon$ for a constant B_1 and any probability measure P where K is a constant. Thus, we can find $\exp\{K/\varepsilon\}$ brackets, $\{[L_j,U_j]\}$, to cover the class $\mathscr{H}_2$ such that $\|U_j-L_j\|_{L_2(P)}\le\varepsilon$ for each pair of

$[L_j, U_j]$. On the other hand, we can further find a partition of $\mathscr{H}_1 = \{\mathbf{h}_1 : \|\mathbf{h}_1\| \leq 1\}$, say $I_1 \bigcup I_2 \bigcup \cdots$, such that the number of partitions is of the order $(1/\varepsilon)$, and, for any $\mathbf{h}_1$ and h_2 in the same partition, their Euclidean distance is less than ε. Therefore, the partition $\{I_1, I_2, \ldots\} \times \{[L_j, U_j]\}$ bracket covers $\mathscr{H}_1 \times \mathscr{H}_2$, and the total number of the partitions is of order $(1/\varepsilon)\exp\{1/\varepsilon\}$. Then, we obtain

$$\log N_{[\cdot]}\big(\varepsilon, \mathscr{G}, L_2(P)\big) \leq O\bigg(\frac{1}{\varepsilon} + \log\varepsilon\bigg).$$

Moreover, $\mathscr{G}$ has an $L_2(P)$-integrable covering function, which is equal to $O(e^{q_1+q_2\|\boldsymbol{Y}\|})$. Hence, from Theorem 2.5.6 (p130) in van der Vaart and Wellner (1996), $\mathscr{G}$ is P-Donsker.

Additionally, from (3.35), we can calculate

$$\begin{aligned}
&\big|l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2]\big| \\
&\leq\ e^{q_1+q_2\|\boldsymbol{Y}\|}\bigg\{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + |\Lambda_s(V_s) - \Lambda_{s0}(V_s)| + \int_0^\tau |\Lambda_s(t) - \Lambda_{s0}(t)|dt\bigg\} \\
&\quad + \bigg|\int_0^\tau \mu_3(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)h_2(t)d(\Lambda_s(t) - \Lambda_{s0}(t)\bigg|. \qquad (3.36)
\end{aligned}$$

If $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \to 0$ and $\sup_{t\in[0,\tau]}|\Lambda_s(t) - \Lambda_{s0}(t)| \to 0$, the above expression converges to zero uniformly. Therefore,

$$\sup_{(\mathbf{h}_1, h_2)\in\mathscr{H}} \mathbf{P}\big[l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2]\big]^2 \longrightarrow 0.$$

3.5.3.3 Derivative operator ∇U_{ξ_0}

From (3.34) in the previous Section 3.5.3.2, we can obtain that

$$\begin{aligned}
&l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2] \\
&= \big[l_\theta(\boldsymbol{\theta}, \Lambda_s)^T - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\big]\mathbf{h}_1 + \big[l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s) - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})\big][h_2]
\end{aligned}$$

$$
\begin{aligned}
&= (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \nabla_{\boldsymbol{\theta}} \mu_1(\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)\mathbf{h}_1 - (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \mu_2(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)\mathbf{h}_1 d\Lambda_s^*(t) \\
&\quad + \int_0^{V_s} \mu_4(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T \mathbf{h}_1 d(\Lambda_s-\Lambda_{s0})(t) \\
&\qquad - \int_0^{V_s}\int_0^{t} \mu_5(u,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T d(\Lambda_s-\Lambda_{s0})(u)\mathbf{h}_1 d\Lambda_s^*(t) \\
&\quad - \int_0^{V_s} \mu_2(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T \mathbf{h}_1 d(\Lambda_s-\Lambda_{s0})(t) \\
&\qquad - (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \mu_3(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) h_2(t) d\Lambda_s^*(t) \\
&\quad - \int_0^{V_s}\int_0^{t} \mu_6(u,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T d(\Lambda_s-\Lambda_{s0})(u) h_2(t) d\Lambda_s^*(t) \\
&\qquad - \int_0^{V_s} \mu_3(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T h_2(t) d(\Lambda_s-\Lambda_{s0})(t) \\
&= (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \left\{ \nabla_{\boldsymbol{\theta}} \mu_1(\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) - \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \mu_2(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) d\Lambda_s^*(t) \right\} \mathbf{h}_1 \\
&\quad + \mathbf{h}_1^T \left\{ \int_0^{\tau} I(t \le V_s) \big[\mu_4(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) - \mu_2(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) \right. \\
&\qquad\qquad \left. - \mu_5(u,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) \int_t^{V_s} d\Lambda_s^*(u) \big] d(\Lambda_s-\Lambda_{s0})(t) \right\} \\
&\quad - (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \int_0^{\tau} I(t \le V_s) \nabla_{\boldsymbol{\theta}} \mu_3(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) h_2(t) d\Lambda_s^*(t) \\
&\quad - \int_0^{\tau} \left\{ I(t \le V_s) \mu_6(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) \int_0^{V_s} h_2(u) d\Lambda_s^*(u) \right. \\
&\qquad\qquad \left. + I(t \le V_s) \mu_3(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*) h_2(t) \right\} d(\Lambda_s-\Lambda_{s0})(t). \qquad (3.37)
\end{aligned}
$$

Then, we have that

$$
\begin{aligned}
&\nabla U_{\xi_0}(\boldsymbol{\theta}-\boldsymbol{\theta}_0, \Lambda_s-\Lambda_{s0})[\mathbf{h}_1, h_2] \\
&= (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \mathbf{P} \left\{ \nabla_{\boldsymbol{\theta}} \mu_1(\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) - \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \mu_2(t,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) d\Lambda_{s0}(t) \right\} \mathbf{h}_1 \\
&\quad + \mathbf{h}_1^T \left\{ \int_0^{\tau} \mathbf{P} \left[I(t \le V_s) \left(\mu_4(t,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) - \mu_2(t,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) \right. \right. \right. \\
&\qquad\qquad \left. \left. \left. - \mu_5(t,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) \int_t^{V_s} d\Lambda_{s0}(u) \right) \right] d(\Lambda_s-\Lambda_{s0})(t) \right\} \\
&\quad - (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \int_0^{\tau} \mathbf{P} \left\{ I(t \le V_s) \nabla_{\boldsymbol{\theta}} \mu_3(t,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0}) \right\} h_2(t) d\Lambda_{s0}(t)
\end{aligned}
$$

$$- \int_0^\tau \mathbf{P}\Big\{ I(t \le V_s)\mu_6(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_0^{V_s} h_2(u) d\Lambda_{s0}(u) + I(t \le V_s)\mu_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) h_2(t) \Big\} d(\Lambda_s - \Lambda_{s0})(t).$$

By the similar algebra done in (3.36), we can verify that, for $j = 1, \ldots, 6$,

$$\sup_{t \in [0,\tau]} ||\mu_j(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) - \mu_j(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})|| \le e^{q_3+q+4||\boldsymbol{Y}||} \Big\{ ||\boldsymbol{\theta}^* - \boldsymbol{\theta}_0|| + \sup_{t \in [0,\tau]} |\Lambda_s^* - \Lambda_{s0}| \Big\},$$

which implies that the linear operator ∇U_{ξ_0} is bounded.

Then, we obtain that

$$\begin{aligned} &\mathbf{P}\big[l_\theta(\boldsymbol{\theta}, \Lambda_s)^T \mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T \mathbf{h}_1 - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2] \big] \\ &= \nabla U_{\xi_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0, \Lambda_s - \Lambda_{s0})[\mathbf{h}_1, h_2] + o\big(||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| + \sup_{t \in [0,\tau]} |\Lambda_s - \Lambda_{s0}| \big)(||\mathbf{h}_1|| + ||h_2||_V). \end{aligned}$$

Therefore, U_ξ is *Fréchet* differentiable at ξ_0.

Additionally, from (3.37) and the above expression, we have

$$\nabla U_{\xi_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0, \Lambda_s - \Lambda_{s0})[\mathbf{h}_1, h_2] = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \Omega_1[\mathbf{h}_1, h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1, h_2] d(\Lambda_s - \Lambda_{s0})(t),$$

where

$$\begin{aligned} \Omega_1[\mathbf{h}_1, h_2] = \; & \mathbf{P}\Big\{ \nabla_{\boldsymbol{\theta}} \mu_1(\boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \mu_2(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\Lambda_{s0}(t) \Big\} \mathbf{h}_1 \\ & - \int_0^\tau \mathbf{P}\Big\{ I(t \le V_s) \nabla_{\boldsymbol{\theta}} \mu_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \Big\} h_2(t) d\Lambda_{s0}(t) \end{aligned}$$

and

$$
\begin{aligned}
&\Omega_2[\mathbf{h}_1, h_2] \\
&= \mathbf{h}_1^T \mathbf{P} \left\{ I(t \le V_s) \big[\mu_4(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \mu_2(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \mu_5(u, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_t^{V_s} d\Lambda_{s0}(u) \big] \right\} \\
&\quad - \mathbf{P} \left\{ I(t \le V_s) \mu_6(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_0^{V_s} h_2(u) d\Lambda_{s0}(u) \right\} \\
&\quad - \mathbf{P} \left\{ I(t \le V_s) \mu_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \right\} h_2(t).
\end{aligned}
$$

Thus, $\Omega = (\Omega_1, \Omega_2)$ is the bounded linear operator from $R^d \times BV[0, \tau]$ to itself. Furthermore, we note that $\Omega = \boldsymbol{H} + (\boldsymbol{M}_1, \boldsymbol{M}_2)$, where

$$
\begin{aligned}
\boldsymbol{H}(\mathbf{h}_1, h_2) &= \big(\mathbf{h}_1, -\mathbf{P} \left\{ I(t \le V_s) \mu_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \right\} h_2(t) \big), \\
\boldsymbol{M}_1(\mathbf{h}_1, h_2) &= \Omega_1[\mathbf{h}_1, h_2] - \mathbf{h}_1, \\
\boldsymbol{M}_2(\mathbf{h}_1, h_2) &= \mathbf{h}_1^T \mathbf{P} \left\{ I(t \le V_s) \big[\mu_4(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \mu_2(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \right. \\
&\qquad \left. - \mu_5(u, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_t^{V_s} d\Lambda_{s0}(u) \big] \right\} \\
&\qquad - \mathbf{P} \left\{ I(t \le V_s) \mu_6(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_0^{V_s} h_2(u) d\Lambda_{s0}(u) \right\},
\end{aligned}
$$

and also note that $\boldsymbol{H}$ is obviously invertible. Since $\boldsymbol{M}_1$ maps into a finite-dimensional space, it is compact. The image of $\boldsymbol{M}_2$ is a continuously differentiable function in $[0, \tau]$. By the Arzela-Ascoli theorem (p41) in van der Vaart and Wellner (1996), $\boldsymbol{M}_2$ is a compact operator from $R^d \times BV[0, \tau]$ to $BV[0, \tau]$. Thus, we conclude that Ω is the summation of an invertible operator $\boldsymbol{H}$ and a compact operator $\boldsymbol{M} = (\boldsymbol{M}_1, \boldsymbol{M}_2)$.

3.6 Simulation Studies

In this section, we present some results from our simulation studies. Two sets of simulations with different generalized linear mixed models for the longitudinal outcomes

are performed. Binary and Poisson data are considered for longitudinal process in the first and second sets of simulations, respectively.

3.6.1 Binary longitudinal outcomes and survival time

In this first set of simulations, we assume Y_{ij} to be a binary outcome following

$$P(Y_{ij} = y_{ij}|b_i) = \exp\left\{y_{ij}\eta_{ij} - \log(1 + \exp\{\eta_{ij}\})\right\}, \qquad y_{ij} = 0,\ 1,$$

with $\eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + b_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3ij} + b_i$ for $j = 1, \ldots, n_i$, and

$$h(t|b_i) = \lambda(t)\exp\{\psi b_i + \boldsymbol{Z}_i(t)\boldsymbol{\gamma}\} = \lambda(t)\exp\{\psi b_i + \gamma_1 Z_{1i} + \gamma_2 Z_{2i}\},$$

where $b_i \sim N(0, \sigma_b^2)$, $X_{1i} \equiv Z_{1i}$ are simulated from a Bernoulli distribution with success probability being 0.5, and $X_{2i} \equiv Z_{2i}$ are simulated from the uniform distribution between 0 and 1. The longitudinal data are generated for every 0.3 unit of time, and thus X_{3ij}, the time at measurement, has the value of every 0.3 unit ranging over 0 through 2.4. We consider different ψ values of -0.1, 0, and 0.1 for negative, zero, and positive dependency between longitudinal process and survival time model, respectively. The parameters in the two models are chosen as $\beta_0 = -1$, $\beta_1 = 1$, $\beta_2 = -0.5$, $\beta_3 = -0.2$, $\sigma_b^2 = 0.5$, $\psi = -0.1/\ 0/\ 0.1$, $\gamma_1 = -0.1$, $\gamma_2 = 0.1$, and $\lambda(t) = 1$. Censoring time is generated from the uniform distribution between 0.4 and 2.4, and the censoring proportion is around 25~35%. We consider different sample sizes (n=200, 400) with 1000 replications. The average number of longitudinal observations (n_i) is 3 with the range of 1 to 8. For the comparison of the estimated baseline cumulative hazards over simulations, we consider the three time points of 0.9, 1.4, 1.9 which are three quartiles of the observed survival time. The results of the maximum likelihood estimates for $\boldsymbol{\theta}$ and baseline cumulative hazards at the given three time points and their respective standard error estimates

are reported in Table 3.3. The simulation study is conducted using R.

In Table 3.3, "True" gives the true values of parameters; the averages of the maximum likelihood estimates from the EM algorithm are in "Est."; the sample standard deviations from 1000 simulations are reported in "SSD"; "ESE" is the average of 1000 standard error estimates based on the observed information matrix; "CP" is the coverage proportion of 95% nominal confidence intervals based on the estimated standard error "ESE". Satterthwaite method is used for the coverage probability of σ_b^2.

From Table 3.3, we can see that even for the smaller sample size (n=200), the bias of the estimates from EM algorithm is negligible for most cases. The estimated standard errors calculated from the observed information matrix are close to the sample standard deviations from the 1000 estimates, and the 95% confidence interval coverage rates are close to 0.95 except those for ψ. The parameter ψ tends to be underestimated with higher than the nominal level coverage rates, but the coverage rate is improved for larger sample size. Thus, with small sample size, the test for ψ is conservative, which strengthens the test results when rejecting the null ($\psi = 0$), and the type I error becomes closer to the nominal level as sample size increases. In addition, the simulations show that the variances of the estimators decrease as the sample size (n) increases. We also can see that the estimates are fairly robust and close to the true values for all different ψ values.

3.6.2 Poisson longitudinal outcomes and survival time

In the second set of simulations, we assume Y_{ij} to follow a Poisson distribution,

$$P(Y_{ij} = y_{ij}|b_i) = \exp\left\{y_{ij}\eta_{ij} - \exp\{\eta_{ij}\} - \log(y_{ij}!)\right\},$$

Table 3.3: Summary of simulation results of maximum likelihood estimation for binary longitudinal outcomes and survival time.

			n=200				n=400			
ψ	Par.	True	Est.	SSD	ESE	CP	Est.	SSD	ESE	CP
- .1	β_0	-1.0	-1.008	.270	.272	.951	-1.006	.194	.191	.945
	β_1	1.0	1.015	.241	.232	.942	.997	.162	.162	.949
	β_2	- .5	- .522	.405	.392	.939	- .500	.282	.274	.945
	β_3	- .2	- .173	.252	.245	.947	- .187	.181	.172	.952
	σ_b^2	.5	.502	.231	.288	.968	.493	.168	.200	.974
	ψ	- .1	- .083	.353	.390	.990	- .102	.235	.245	.978
	γ_1	- .1	- .102	.174	.174	.949	- .104	.121	.121	.955
	γ_2	.1	.099	.296	.303	.953	.097	.214	.210	.950
	$\Lambda(.9)$	.9	.910	.181	.184	.955	.906	.132	.128	.943
	$\Lambda(1.4)$	1.4	1.444	.294	.299	.962	1.421	.211	.204	.943
	$\Lambda(1.9)$	1.9	1.980	.446	.449	.955	1.945	.312	.302	.951
0	β_0	-1.0	-1.012	.281	.273	.944	-1.008	.192	.192	.948
	β_1	1.0	1.014	.235	.233	.954	1.006	.164	.163	.948
	β_2	- .5	- .501	.414	.393	.934	- .503	.278	.276	.949
	β_3	- .2	- .190	.263	.246	.936	- .194	.175	.173	.950
	σ_b^2	.5	.505	.237	.290	.960	.503	.176	.203	.963
	ψ	.0	.008	.375	.388	.994	.003	.236	.241	.980
	γ_1	- .1	- .108	.180	.174	.942	- .103	.113	.121	.968
	γ_2	.1	.098	.309	.303	.949	.101	.209	.210	.951
	$\Lambda(.9)$	.9	.920	.188	.186	.952	.905	.131	.127	.944
	$\Lambda(1.4)$	1.4	1.463	.306	.303	.948	1.415	.206	.202	.952
	$\Lambda(1.9)$	1.9	2.006	.462	.457	.953	1.937	.306	.299	.948
.1	β_0	-1.0	-1.009	.285	.274	.948	-1.000	.192	.192	.945
	β_1	1.0	1.004	.224	.234	.964	1.004	.166	.163	.952
	β_2	- .5	- .510	.414	.395	.945	- .512	.284	.276	.943
	β_3	- .2	- .186	.260	.249	.948	- .185	.189	.175	.929
	σ_b^2	.5	.519	.249	.295	.946	.498	.174	.203	.965
	ψ	.1	.117	.354	.386	.990	.129	.246	.247	.986
	γ_1	- .1	- .096	.175	.175	.946	- .101	.117	.122	.966
	γ_2	.1	.086	.312	.304	.944	.101	.212	.211	.948
	$\Lambda(.9)$	.9	.915	.184	.185	.957	.904	.129	.127	.946
	$\Lambda(1.4)$	1.4	1.455	.305	.303	.955	1.413	.203	.203	.954
	$\Lambda(1.9)$	1.9	2.010	.481	.463	.952	1.938	.303	.302	.959

with η_{ij} defined as in Section 3.6.1. We also consider the same hazards model and simulation setting as those used in Section 3.6.1 except $\sigma_b^2 = 0.2$. The simulated Poisson longitudinal outcomes range over 0 to 7 with the average 0.5.

Table 3.4 shows that overall the estimates perform well even for the smaller sample size $n = 200$ with small biases of the estimates except ψ. We conducted additional simulations with sample sizes of 800 and 1000, and the bias of ψ existing for the small sample size decreases as sample size increases over 200, 400, 800 and 1000. The estimated standard errors using the observed information matrix are close to the sample standard deviations, and the 95% confidence interval coverage rates are close to 0.95 except for σ_b^2 and ψ.

From Table 3.4, ψ is seemingly underestimated with higher than the nominal coverage rates, but the coverage rate decreases to close to 95% nominal level as sample size increases. Additional simulations we conducted show that, with sample sizes of 800, the 95% confidence interval coverage rates for ψ =-0.1, 0 and 0.1 were 95.5%, 95.9% and 95.9%, respectively. σ_b^2 also appears to have high coverage rates, which may be due to numerical problem since its coverage rates are still high for larger sample sizes. This implies that variance of σ_b^2 may not be estimated well for Poisson longitudinal distribution. In the meantime, the test for σ_b^2 is conservative, which strengthens the test result for rejecting the null $(\sigma_b^2 = 0)$. On the other hand, profile likelihood may be an alternative estimation approach for σ_b^2. It is also shown that the variances of the estimators decrease for larger sample size, and the estimates are fairly robust and close to the true values for all three different ψ values.

3.7 Analysis of the CHANCE Study

We now return to the CHANCE study described in Section 3.2, and apply our proposed method to Head and Neck Cancer Specific symptoms (HNCS) among QoL domains with

Table 3.4: Summary of simulation results of maximum likelihood estimation for Poisson longitudinal outcomes and survival time.

			n=200				n=400			
ψ	Par.	True	Est.	SSD	ESE	CP	Est.	SSD	ESE	CP
- .1	β_0	-1.0	-1.001	.186	.192	.959	-1.005	.135	.135	.949
	β_1	1.0	1.014	.165	.161	.949	1.008	.118	.113	.933
	β_2	- .5	- .513	.275	.260	.938	- .502	.189	.182	.951
	β_3	- .2	- .188	.178	.164	.941	- .189	.128	.115	.921
	σ_b^2	.2	.195	.074	.096	.978	.197	.051	.067	.986
	ψ	- .1	- .056	.603	.621	.981	- .075	.417	.388	.970
	γ_1	- .1	- .098	.177	.175	.952	- .098	.126	.121	.940
	γ_2	.1	.081	.311	.305	.946	.103	.218	.211	.956
	$\Lambda(\ .9)$	.9	.922	.189	.187	.949	.902	.130	.127	.946
	$\Lambda(1.4)$	1.4	1.466	.321	.307	.941	1.418	.209	.204	.950
	$\Lambda(1.9)$	1.9	2.036	.502	.474	.950	1.947	.308	.304	.950
0	β_0	-1.0	-1.000	.190	.192	.946	-1.003	.134	.135	.948
	β_1	1.0	1.009	.162	.161	.944	1.004	.114	.113	.949
	β_2	- .5	- .512	.277	.260	.933	- .496	.187	.183	.940
	β_3	- .2	- .190	.179	.165	.934	- .187	.125	.116	.935
	σ_b^2	.2	.195	.076	.096	.977	.199	.052	.067	.984
	ψ	.0	.017	.606	.628	.989	.061	.403	.382	.968
	γ_1	- .1	- .098	.174	.175	.952	- .105	.125	.121	.940
	γ_2	.1	.094	.311	.305	.952	.101	.212	.211	.950
	$\Lambda(\ .9)$	.9	.914	.185	.185	.955	.906	.130	.128	.944
	$\Lambda(1.4)$	1.4	1.459	.305	.304	.954	1.428	.209	.205	.942
	$\Lambda(1.9)$	1.9	2.016	.472	.464	.957	1.961	.308	.305	.945
.1	β_0	-1.0	-1.000	.192	.193	.947	-1.005	.131	.135	.958
	β_1	1.0	1.013	.168	.162	.939	1.007	.115	.114	.945
	β_2	- .5	- .522	.279	.261	.937	- .500	.191	.183	.939
	β_3	- .2	- .193	.191	.167	.929	- .188	.129	.117	.928
	σ_b^2	.2	.198	.076	.096	.978	.197	.053	.067	.984
	ψ	.1	.152	.609	.627	.993	.137	.403	.390	.975
	γ_1	- .1	- .098	.174	.176	.944	- .103	.120	.121	.951
	γ_2	.1	.089	.303	.306	.953	.090	.203	.211	.958
	$\Lambda(\ .9)$	.9	.919	.187	.187	.945	.913	.122	.128	.962
	$\Lambda(1.4)$	1.4	1.461	.317	.308	.952	1.439	.205	.207	.961
	$\Lambda(1.9)$	1.9	2.038	.519	.480	.951	1.965	.308	.307	.956

survival time in this section. We are interested in testing the correlation between survival time and longitudinal QoL outcomes and investigating which variables predict the QoL satisfaction or the risk of death or both. In the full models for both longitudinal QoL and survival time, race, the number of 12 oz. beers consumed per week, household income, surgery, radiation therapy, chemotherapy, tumor site, and tumor stage are considered as categorical, and age at diagnosis, the number of persons supported by household income, body mass index (BMI), and the total number of medical conditions reported as continuous. Additionally, 2 interactions with race, i.e. race × the total number of medical conditions reported and race × tumor site, are included in both models since we are particularly interested in the difference of QoL and survival between African American and White. Time at survey measurement is also included as a covariate for longitudinal outcomes. A random intercept for the dependence between the QoL satisfaction and the risk of death is included in both models, and assumed to follow a normal distribution with mean zero. In addition to the simultaneous analysis, we also conduct separate analyses fitting the generalized linear mixed model and the Cox proportional hazards model to the longitudinal QoL and survival time respectively and compare the results to those from our proposed simultaneous method.

After fitting the simultaneous models with all the covariates, we use backward variable selection based on the Likelihood Ratio Test (LRT) and find that surgery, chemotherapy, tumor site, age at diagnosis, and all 2 interactions are not statistically significant in both models for HNCS QoL satisfaction and survival time at the significance level 0.05. We remove these variables and refit the simultaneous models. Then, the LRT shows that race, radiation therapy, the number of persons supported by household income, BMI, and the total number of medical conditions reported are not statistically significant for the risk of death. We further reduce the models by removing them from the hazards model and refit the reduced simultaneous models. Table 3.5

gives the results from this final models. From the "Simultaneous" columns, we see that the number of 12 oz. beers consumed per week, household income, tumor stage, and the total number of medical conditions reported are significantly associated with both patients' HNCS QoL satisfaction and hazard of death. Using 30 or more of 12 oz. beers consumed per week as the reference group, all categories of the smaller amount are associated with HNCS QoL satisfaction and lower risk of death, higher household income is overall associated with HNCS QoL satisfaction and lower risk of death, and both patients' HNCS QoL satisfaction and risk of death are significantly different for patients in different tumor stages. Specifically, for instance, with the log-scaled odds and hazard ratios of 1.060 and -1.076 for HNCS QoL satisfaction and death respectively, patients who consumed 5 to 14 of 12 oz. beers per week appear to have 2.886 times odds for HNCS QoL satisfaction and 0.341 times hazards of death compared to those that consumed 30 or more of 12 oz. beers per week in the study after adjusting for the other covariates in the model. Looking at the number of medical conditions reported, for each additional medial condition reported, the odds ratio of HNCS QoL satisfaction is decreased by 16% and the hazard of death is increased by 29%. That is, patients with a greater number of medical conditions reported have lower HNCS QoL satisfaction and higher risk of death after adjusting for the other covariates in the model. In the meantime, race (African-American), radiation therapy, the number of persons supported by household income, and BMI are selected only in the HNCS QoL longitudinal model. African-Americans, patients not treated with radiation therapy, patients in the family with the smaller number of persons supported by household income, or patients with higher BMI are likely to be more satisfied with longitudinal HNCS QoL while the risk of death is not affected by race, radiation therapy, the number of persons supported by household income and BMI. Furthermore, we also find that time at survey measurement is statistically significant in the HNCS QoL longitudinal model implying that

patients are more satisfied over time. The parameter ψ for the dependence between longitudinal HNCS QoL and survival time is negative and has p-value as 0.131. This implies that the longitudinal HNCS QoL and survival time are marginally correlated and some latent factors which increase HNCS QoL satisfaction also decrease the risk of death. For the purpose of comparison, we conducted separate analyses for longitudinal HNCS QoL and survival time whose results are given in the last three columns of Table 3.5. The generalized linear mixed model (GLMM) and the Cox proportional hazards model are used for longitudinal outcomes and survival time respectively. The GLMM also considers individual heterogeneity through subject-specific random effects although it does not incorporate the correlation between longitudinal outcomes and survival time. Comparing the results from the simultaneous and separate analyses of Table 3.5, we can see our simultaneous analysis additionally indicates the number of persons supported by household income, BMI, and the total number of medical conditions reported in the HNCS QoL longitudinal model (p-values=0.025, 0.007, and 0.030, respectively) and the number of 12 oz. beers consumed per week in the hazard model (p-values=0.045 and 0.005 for 'None' and '5 to 14') as significant which are not selected by separate analyses. Figure 3.1 shows the estimated baseline cumulative hazard rates over follow-up time with the 95% confidence interval. Since the baseline cumulative hazard rates are bounded by 0, we first log-transformed the estimated baseline cumulative hazard rates and obtained the 95% lower and upper bounds for the log-scaled estimated baseline cumulative hazards. Then, we re-transformed them into their original scale. The estimated baseline cumulative hazard rates look flat at the very early time within a year, but soon appear to be linearly increasing. Figure 3.2 shows the Kaplan-Meier estimates (solid line) and the predicted survival probabilities based on the simultaneous analysis (dashed line). These two survival curves are very close to each other which implies our proposed method fits the data well.

Table 3.5: Analyses results for the HNCS QoL and survival time of the CHANCE study

Parameter		Simultaneous Est.	Simultaneous ESE	Simultaneous P-value	Separate Est.	Separate ESE	Separate P-value
< HNCS QoL longitudinal model >							
Intercept	β_0	.744	.538	.167	1.190	.390	.002
Race (ref= White)							
– African American	β_1	.564	.229	.014	.511	.256	.047
# of 12 oz. beers consumed per week (ref=30 or more)							
– None	β_2	.636	.269	.018	.622	.300	.038
– less than 1	β_3	.830	.357	.020	.735	.396	.064
– 1 to 4	β_4	1.302	.294	<.001	1.268	.326	<.001
– 5 to 14	β_5	1.060	.251	<.001	1.018	.279	<.001
– 15 to 29	β_6	.601	.289	.037	.547	.327	.095
Household income (ref= level1: 0–10K)							
– level2: 20–30K	β_7	-.271	.231	.241	-.328	.258	.204
– level3: 40–50K	β_8	.297	.255	.245	.250	.282	.376
– level4: ≥ 60K	β_9	1.199	.274	<.001	1.045	.286	<.001
Radiation therapy (ref= No)							
– Yes	β_{10}	-1.132	.260	<.001	-1.048	.280	<.001
Tumor stage (ref= I)							
– II	β_{11}	-.416	.300	.166	-.352	.330	.286
– III	β_{12}	-1.335	.284	<.001	-1.198	.314	<.001
– IV	β_{13}	-1.175	.254	<.001	-1.057	.277	<.001
# of persons supported by household income	β_{14}	-.189	.084	.025			
BMI	β_{15}	.041	.015	.007			
Total # of medical conditions reported	β_{16}	-.175	.080	.030			
Time at survey measurement (years)	β_{17}	.241	.066	<.001	.254	.067	<.001
variance of random effects	σ_b^2	.303	.173	.013	1.169	.257	
< Hazards model >							
Random effect coefficient	ψ	-1.427	.946	.131			
# of 12 oz. beers consumed per week (ref=3 or more)							
– None	γ_1	-.772	.386	.045			
– less than 1	γ_2	-.155	.426	.715			
– 1 to 4	γ_3	-.802	.414	.053			
– 5 to 14	γ_4	-1.076	.383	.005			
– 15 to 29	γ_5	-.591	.399	.139			
Household income (ref= level1: 0–10K)							
– level2: 20–30K	γ_6	-.218	.294	.459	-.219	.263	.406
– level3: 40–50K	γ_7	-.941	.371	.011	-.928	.331	.005
– level4: ≥ 60K	γ_8	-1.463	.401	<.001	-1.393	.358	<.001
Tumor stage (ref= I)							
– II	γ_9	-.199	.465	.668	-.295	.435	.498
– III	γ_{10}	.235	.433	.588	.136	.389	.727
– IV	γ_{11}	1.059	.360	.003	.914	.295	.002
Total # of medical conditions reported	γ_{12}	.256	.110	.020	.205	.091	.025

P-value for testing σ_b^2 being zero is based on a mixture of 0 and χ^2 distribution with 1 degree of freedom with equal mixing probabilities.

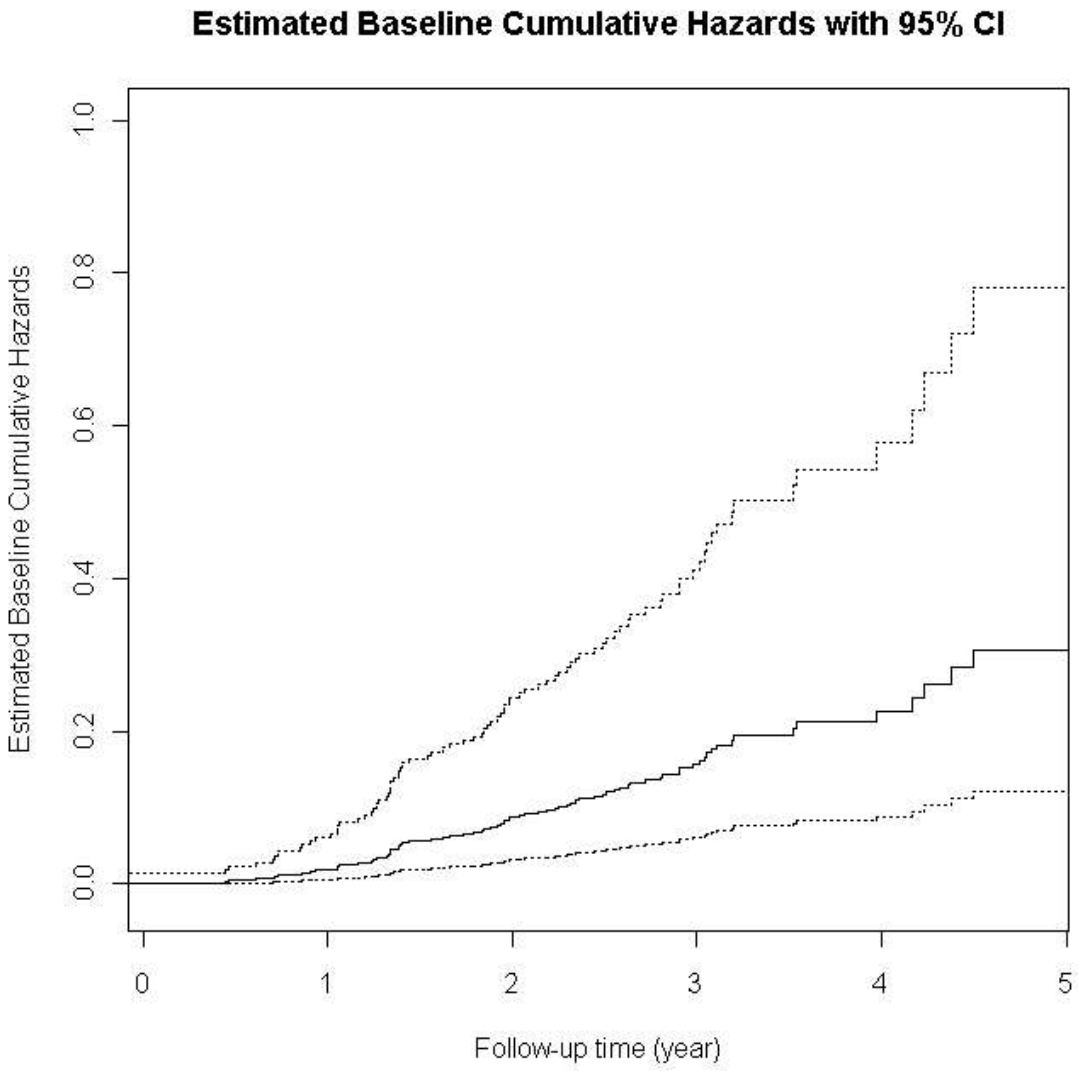

Figure 3.1: Estimated baseline cumulative hazards (solid line) with 95% confidence interval (dashed lines) by the simultaneous analysis of HNCS QoL longitudinal outcome and survival time

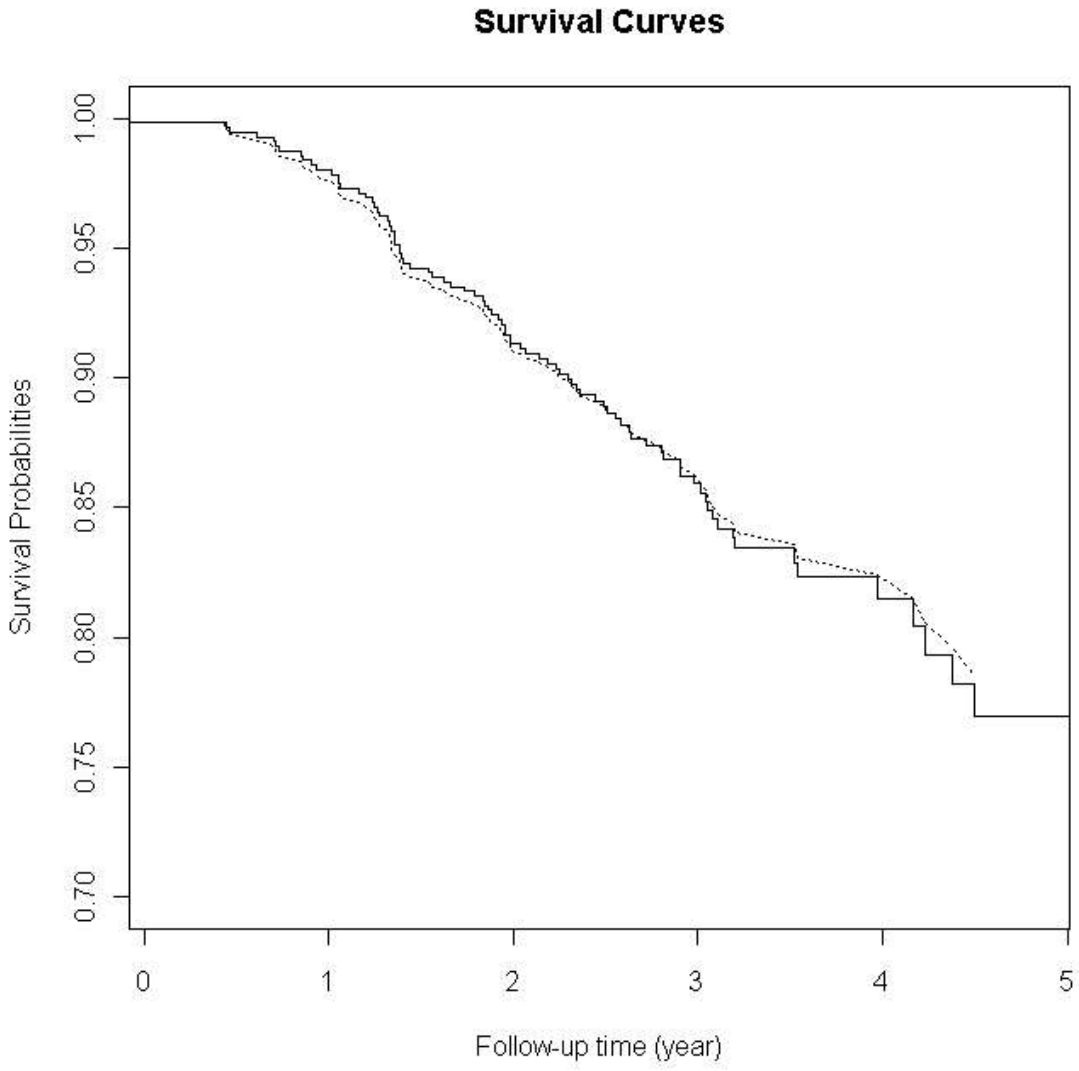

Figure 3.2: Kaplan-Meier estimates (solid line) and the predicted survival probabilities based on the simultaneous analysis of HNCS QoL longitudinal outcome and survival time (dashed line)

3.8 Concluding Remarks

We have proposed a method for the simultaneous modeling of longitudinal outcomes including both categorical and continuous data with a generalized linear mixed model and survival time with a stratified multiplicative proportional hazards model through random effects. We have also developed a maximum likelihood estimation method for the proposed simultaneous model, and presented asymptotic properties of the proposed estimators. The proposed estimation procedure using EM algorithm has been assessed via simulation studies. The proposed estimates performed well in finite samples. The variance estimates based on the observed information matrix approximate the true variance well in finite samples.

The proposed method was applied to the CHANCE study data. The results for longitudinal HNCS and survival time have shown that, after adjusting for the other covariates in the simultaneous models, the lower amount of beers consumed per week, higher household income, lower stage, and the lower total number of medical conditions reported are associated with more HNCS QoL satisfaction and lower risk of death. Further, African-Americans, patients not treated with radiation therapy, patients in the family with the smaller number of persons supported by household income, or patients with higher BMI are likely to be more satisfied with longitudinal HNCS QoL while the risk of death is not affected by race, radiation therapy, the number of persons supported by household income and BMI. Time at survey measurement in the HNCS QoL longitudinal model is also statistically significant implying that patients are more satisfied over time. Furthermore, our proposed method additionally finds more predictors including: the number of persons supported by household income, BMI, and the total number of medical conditions reported in the HNCS QoL longitudinal model and the predictor, number of 12 oz. beers consumed per week, in the hazard model while separate analyses do not select them. This result comparing the simultaneous and separate

analyses supports that, even when the longitudinal outcomes and survival time are only marginally correlated, our simultaneous analysis could provide better power than separate analyses not considering the dependency between the longitudinal outcomes and survival time.

In our proposed method, all the information on survival, longitudinal outcomes, and covaraties are used. As a result of this, the parameter estimates can be more efficient. The proposed model also generalizes previous work to general longitudinal outcomes. This work fills in some gaps in the joint modeling research. Future work can include relaxing normal assumption for the random effects and considering generalization to mixed types of longitudinal outcomes.

Chapter 4

JOINT MODELING OF SURVIVAL TIME AND LONGITUDINAL OUTCOMES WITH FLEXIBLE RANDOM EFFECTS

4.1 Introduction

In biomedical or public health research, it is common that both longitudinal outcomes over time and survival endpoint are collected for the same subject along with the subject's characteristics or risk factors. Investigators are interested in finding important variables which predict both longitudinal outcomes and survival time. Among the existing approaches for longitudinal data and survival time, the selection model and the pattern mixture model have been widely used. The selection model estimating the distribution of survival time given longitudinal data was studied by numerous authors, for example, Tsiatis *et al.* (1995), Tsiatis and Davidian (2001), Xu and Zeger (2001a,b) and Tseng *et al.* (2005). The pattern mixture model focuses on the trend of longitudinal outcomes conditional on survival time and was studied by Wu and Carroll (1988), Hogan and Laird (1997), Albert and Follmann (2000, 2007) and Ding and Wang (2008) among others. On the other hand, simultaneous modeling of the longitudinal and survival data

was proposed by Xu and Zeger (2001b), Zeng and Cai (2005), Elashfoff *et al.* (2007, 2008) and Rizopoulos *et al.* (2008). Wang and Taylor (2001), Brown and Ibrahim (2003) and Hu *et al.* (2009) studied simultaneous modeling in the Bayesian perspective.

In all the joint models, random effects are incorporated to accommodate the latent dependence between survival time and longitudinal outcomes. Random effects are conventionally assumed to be normally distributed. However, it is unclear whether the normality assumption is truly satisfied in practice. Furthermore, misspecifying normality assumption can lead to serious bias in estimation (Neuhaus *et al.*, 1992; Kleinman and Ibrahim, 1998; Heagerty and Kurland, 2001; Agresti *et al.*, 2004).

In this paper, we assume that the underlying distribution of random effects is unknown. In estimating model parameters, we propose to use a mixture of Gaussian distributions as an approximation for the unknown random effect distribution. Moreover, we simultaneously model the survival time with a stratified Cox proportional hazards model and longitudinal outcomes with a generalized linear mixed model to incorporate both categorical and continuous longitudinal outcomes. Finite sample properties of the proposed estimators and robustness of the mixture distribution are assessed via simulations. We adopt AIC and BIC for selecting the number of mixtures and also conduct simulation studies to assess these selection procedures.

The outline of this paper is as follows. In Section 4.2, we present a simultaneous modeling for longitudinal outcomes and survival time with random effects from an unknown distribution, and describe the inference procedure. Asymptotic properties of the proposed estimators and the technical details of their proofs are investigated in Section 4.3 and Section 4.4, respectively. Numerical results from simulation studies are given in Section 4.5. Our proposed method is illustrated with the data from the Carolina Head and Neck Cancer Study (CHANCE) in Section 4.6. In Section 4.7, we discuss some further consideration. EM-algorithms are provided in Appendix.

4.2 Models and Inference Procedure

4.2.1 Model formulation and notation

We use $Y(t)$ to denote the value of a longitudinal marker process at time t. Suppose $Y(t)$ is from a distribution belonging to exponential family in order to incorporate both continuous and categorical measurements. Let T denote survival time, and suppose that the survival time T is possibly right censored. Suppose a set of n subjects are followed over an interval $[0, \tau]$, where τ is the study end time. Denote $\boldsymbol{b}_i^*$, $i = 1, \ldots, n$, as a vector of subject-specific random effects of dimension d_b and $\boldsymbol{b}_i^*$'s are mutually independent.

Given the random effects $\boldsymbol{b}_i^*$, the observed covariates, and the observed outcome history till time t, we assume that the longitudinal outcome $Y_i(t)$ at time t for subject i follows a distribution from the exponential family with density,

$$\exp\left\{\frac{y_i\eta_i(t) - B(\eta_i(t))}{A(D_i(t;\phi))} + C(y_i, D_i(t;\phi))\right\} \tag{4.1}$$

with $\mu_i(t) = \mathrm{E}(Y_i(t)|\boldsymbol{b}_i^*) = B'(\eta_i(t))$ and $v_i(t) = \mathrm{Var}(Y_i(t)|\boldsymbol{b}_i^*) = B''(\eta_i(t))A(D_i(t;\phi))$, satisfying

$$\eta_i(t) = g(\mu_i(t)) = \boldsymbol{X}_i(t)\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_i(t)\boldsymbol{b}_i^*$$

and $v_i(t) = v(\mu_i(t))A(D_i(t;\phi))$, where $g(\cdot)$ and $v(\cdot)$ are known link and variance functions respectively, $\boldsymbol{X}_i(t)$ and $\widetilde{\boldsymbol{X}}_i(t)$ are the row vectors of the observed covariates for subject i, and $\boldsymbol{\beta}$ is a column vector of coefficients for $\boldsymbol{X}_i(t)$. $\boldsymbol{X}_i(t)$ does not include intercept and it does not contain any covariates in $\widetilde{\boldsymbol{X}}_i(t)$ because the intercept and any potential common covariates for fixed effects are combined with the corresponding random effects in $\widetilde{\boldsymbol{X}}_i(t)$ so that mean of random effects does not have restriction.

Given the random effects $\boldsymbol{b}_i^*$, the observed covariates, and the observed survival history before time t, the conditional hazard rate function for the survival time T_i of

subject i is assumed to follow a stratified multiplicative hazards model,

$$\lambda_s(t)\exp\{\widetilde{\boldsymbol{Z}}_i(t)(\boldsymbol{\psi}\circ\boldsymbol{b}_i^*)+\boldsymbol{Z}_i(t)\boldsymbol{\gamma}\}, \tag{4.2}$$

where $\boldsymbol{Z}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ are the row vectors of the observed covariates and may share some components, $\boldsymbol{\psi}$ is a vector of parameters of the coefficients for random effects, $\lambda_s(t)$ is the s-th stratum baseline hazard rate function, and $\boldsymbol{\gamma}$ is a column vector of coefficients for $\boldsymbol{Z}_i(t)$. Note that $\boldsymbol{Z}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ do not include dummy variables for strata since baseline hazard rate is stratum-specific. Here, for any vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ of the same dimension, $\boldsymbol{a}_1\circ\boldsymbol{a}_2$ denotes the component-wise product. In addition, $\widetilde{\boldsymbol{X}}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ have the same dimensions as $\boldsymbol{b}_i^*$'s. For the subject-specific random effects $\boldsymbol{b}_i^*$, we assume the underlying distribution of $\boldsymbol{b}_i^*$ is unknown and denote its density as $f(\boldsymbol{b}_i^*)$.

Let n_i be the number of the observed longitudinal measurements for subject i, and assume that the distributions of n_i and the observation times for longitudinal measurements are independent of the parameters of interest in this joint model. The observed data from n subjects are $(n_i, Y_{ij}, \boldsymbol{X}_{ij}, \widetilde{\boldsymbol{X}}_{ij})$, $j=1,\ldots,n_i$, $i=1,\ldots,n$, and $(V_i, \Delta_i, S_i, \{(\boldsymbol{Z}_i(t), \widetilde{\boldsymbol{Z}}_i(t)) : t\le V_i\})$, i=1,…,n, where for subject i, $(Y_{ij}, \boldsymbol{X}_{ij}, \widetilde{\boldsymbol{X}}_{ij})$ is the j-th observation of $(Y_i(t), \boldsymbol{X}_i(t), \widetilde{\boldsymbol{X}}_i(t))$, C_i is the right-censoring time and independent of T_i and $Y_i(t)$ given the covariates and the random effects, $V_i = \min(T_i, C_i)$, S_i denotes the stratum, and $\Delta_i = I(T_i \le C_i)$. For all n subjects, we write $\boldsymbol{Y} = (\boldsymbol{Y}_1^T,\ldots,\boldsymbol{Y}_n^T)^T$, $\boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{in_i})^T$, $\boldsymbol{V} = (V_1,\ldots,V_n)^T$, and $\boldsymbol{b}^* = (\boldsymbol{b}_1^{*T},\ldots,\boldsymbol{b}_n^{*T})^T$. Then, the likelihood function of the complete data $(\boldsymbol{Y},\boldsymbol{V},\boldsymbol{b}^*)$ has the form,

$$\begin{aligned} L_c(\boldsymbol{Y},\boldsymbol{V},\boldsymbol{b}^*) &= \prod_{i=1}^{n} f(\boldsymbol{Y}_i|\boldsymbol{b}_i^*)\left(\prod_{s=1}^{S}\left[f(V_i|\boldsymbol{b}_i^*)\right]^{I(S_i=s)}\right)f(\boldsymbol{b}_i^*) \\ &= \prod_{i=1}^{n}\exp\left\{\sum_{j=1}^{n_i}\left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i^*)-B(\boldsymbol{\beta};\boldsymbol{b}_i^*)}{A(D_i(t_j;\phi))}+C(Y_{ij};D_i(t_j;\phi))\right]\right\} \end{aligned}$$

$$\times \left(\prod_{s=1}^{S} \left[\lambda_s(V_i)^{\Delta_i} \exp\left\{ \Delta_i \big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_i^*) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma} \big] \right.\right.\right.$$
$$\left.\left.\left. - \int_0^{V_i} \exp\left\{ \widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_i^*) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma} \right\} d\Lambda_s(u) \right\} \right]^{I(S_i=s)} \right)$$
$$\times \ f(\boldsymbol{b}_i^*), \tag{4.3}$$

and the full likelihood function of the observed data $(\boldsymbol{Y}, \boldsymbol{V})$ is expressed as

$$L_f(\boldsymbol{Y}, \boldsymbol{V}) \ = \ \int_{\boldsymbol{b}^*} L_c(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}^*) d\boldsymbol{b}^*. \tag{4.4}$$

The parameter $\boldsymbol{\psi}$ in model (4.2) characterizes the dependence between the longitudinal outcomes and the survival time due to latent random effects: $\boldsymbol{\psi} = \mathbf{0}$ means the dependence between the survival time and longitudinal responses are not due to these latent variables; $\boldsymbol{\psi} \neq \mathbf{0}$ means such dependence may be due to these latent variables. In other words, $\boldsymbol{\psi} > \mathbf{0}$ implies there may be some latent factors increasing both the longitudinal outcomes and the risk of survival endpoint simultaneously while $\boldsymbol{\psi} < \mathbf{0}$ implies some latent factors causing the increment of longitudinal outcomes may decrease the risk of survival endpoint.

4.2.2 Inference procedure

For parameter estimation, we approximate the random effect distribution with a mixture of Gaussian distributions. This method was studied in some literature to extend normality assumption of random effects. For instance, Verbeke and Lesaffre (1996), Verbeke and Molenberghs (2000), and Zhang and Davidian (2001) used it in a linear mixed model, and Verbeke and Lesaffre (1996), Fieuws *et al.*. (2004), and Caffo *et al.* (2007) considered it in a GLMM. Alternatively, Ghidey *et al.*. (2004) and Komárek and Lesaffre (2008a) used the penalized Gaussian mixture (PGM) approach in a linear mixed model and a GLMM respectively. Furthermore, Komárek and Lesaffre (2008b,

2009) suggested a Bayesian accelerated failure time (AFT) model with random effects following a PGM.

For the subject-specific random effects $\boldsymbol{b}_i^*$ in Section 4.2.1, we approximate the distribution of $\boldsymbol{b}_i^*$ with a mixture of a finite number of d_b-dimensional multivariate normal distributions. That is, the distribution of $\boldsymbol{b}_i^*$ is approximated by $\sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_b)$, where K is the number of mixture components. We denote the probability of belonging to component k by w_k, such that $\sum_{k=1}^{K} w_k = 1$. $\boldsymbol{\mu}_k$ is the mean of the k-th component and it is assumed that each component has the same covariance matrix $\boldsymbol{\Sigma}_b$. This constraint is needed to avoid infinite likelihoods (Böhning, 1999). We write $\boldsymbol{w} = (w_1, \ldots, w_{K-1})^T$, the vector of $K-1$ component probabilities, and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_K^T)^T$, the vector of all component means. We introduce $\boldsymbol{b}_i$ and $\alpha_i = k$, $(k = 1, \ldots, K)$, as the i-th subject's random effects following the mixture distribution and the k-th component of the mixture from which $\boldsymbol{b}_i$ is sampled, respectively. The distribution of α_i is then described by $P(\alpha_i = k) = w_k$ and, given $\alpha_i = k$, $\boldsymbol{b}_{ik} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_b)$. Thus, $\boldsymbol{b}_i = \sum_{k=1}^{K} I(\alpha_i = k)\boldsymbol{b}_{ik}$, where $I(\alpha_i = k)$ is the indicator of belonging to component k. For n subjects, $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T)^T$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$.

Now we estimate and make inferences on the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\phi}^T, \mathrm{Vec}(\boldsymbol{\Sigma}_b)^T, \boldsymbol{\mu}^T, \boldsymbol{w}^T, \boldsymbol{\psi}^T, \boldsymbol{\gamma}^T)^T$ and the baseline cumulative hazard functions with S strata, $\boldsymbol{\Lambda}(t) = (\Lambda_1(t), \ldots, \Lambda_S(t))^T$, where $\Lambda_s(t) = \int_0^t \lambda_s(u)du$, $s = 1, \ldots, S$. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are from the longitudinal model, $\boldsymbol{\psi}$ and $\boldsymbol{\gamma}$ are from the hazard model, and, $\boldsymbol{\mu}$, $\boldsymbol{w}$, and $\boldsymbol{\Sigma}_b$ are associated with the random effects $\mathrm{Vec}(\cdot)$ operator creates a column vector from a matrix by stacking the diagonal and upper-triangle elements of the matrix. The likelihood function (4.3) of the complete data $(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}, \boldsymbol{\alpha})$ and the full likelihood function (4.4) of the observed data $(\boldsymbol{Y}, \boldsymbol{V})$ for $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ have the following forms respectively,

$$
\begin{aligned}
&L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}, \boldsymbol{\alpha}) \\
&= \prod_{i=1}^{n} \prod_{k=1}^{K} \Bigg[\exp\left\{ \sum_{j=1}^{n_i} \left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_{ik}) - B(\boldsymbol{\beta}; \boldsymbol{b}_{ik})}{A(D_i(t_j;\phi))} + C(Y_{ij}; D_i(t_j;\phi)) \right] \right\} \\
&\qquad \times \Bigg(\prod_{s=1}^{S} \Bigg[\lambda_s(V_i)^{\Delta_i} \exp\Big\{ \Delta_i \big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_{ik}) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma} \big] \\
&\qquad\qquad - \int_0^{V_i} \exp\big\{ \widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_{ik}) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma} \big\} d\Lambda_s(u) \Big\} \Bigg]^{I(S_i = s)} \Bigg) \\
&\qquad \times (2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{b}_{ik} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_b^{-1} (\boldsymbol{b}_{ik} - \boldsymbol{\mu}_k) \right\} \times w_k \Bigg]^{I(\alpha_i = k)}
\end{aligned}
$$

and $\quad L_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) = \sum_{\boldsymbol{\alpha}} \int_{\boldsymbol{b}} L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}, \boldsymbol{\alpha}) d\boldsymbol{b}.$

The proposed estimation method is to calculate the maximum likelihood estimates for $(\boldsymbol{\theta}, \boldsymbol{\Lambda}(t))$ over a set of $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}(t)$. We let each $\Lambda_s(t)$ of $\boldsymbol{\Lambda}(t)$, $s = 1, \ldots, S$, be a non-decreasing and right-continuous step function with jumps only at the observed failure times belonging to stratum s.

EM-algorithm is used for calculating the maximum likelihood estimates. In the EM-algorithm, $\boldsymbol{b}_i$ and α_i are considered as missing data for $i = 1, \ldots, n$. Therefore, the M-step solves the conditional score equations from complete data given observations, where the conditional expectation can be evaluated in E-step. The procedure involves iterating between the following two steps until convergence is achieved: at the m-th iteration,

(1) E-step Calculate the conditional expectations of some known functions of $\boldsymbol{b}_i$ and α_i, needed in the next M-step, for subject i with $S_i = s$ given observations and the current estimate $(\boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)})$. The conditional expectation is calculated using the Gauss-Hermite Quadrature numerical approximation, denoted as $E[q(\boldsymbol{b}_i, \alpha_i) | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}]$ for a

known function $q(\boldsymbol{b}_i, \alpha_i)$.

(2) M-step After differentiating the conditional expectation of complete data log-likelihood function given observations and the current estimate $(\boldsymbol{\theta}^{(m)}, \boldsymbol{\Lambda}^{(m)})$, the updated estimator $(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\Lambda}^{(m+1)})$ can be obtained as follows: $(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\phi}^{(m+1)})$ solves the conditional expectation of complete data log-likelihood score equation using one-step Newton-Raphson iteration; For the covariance matrix of random effects,

$$\boldsymbol{\Sigma}_b^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\sum_{k=1}^{K} \mathrm{E}\left[I(\alpha_i = k)(\boldsymbol{b}_{ik} - \boldsymbol{\mu}_k)(\boldsymbol{b}_{ik} - \boldsymbol{\mu}_k)^T | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}\right] I(S_i = s);$$

For the k-th mixture component $(k = 1, \ldots, K)$,

$$\mu_k^{(m+1)} = \frac{\sum_{i=1}^{n}\sum_{s=1}^{S} \mathrm{E}\left[I(\alpha_i = k)\boldsymbol{b}_{ik} | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}\right] I(S_i = s)}{\sum_{i=1}^{n}\sum_{s=1}^{S} \mathrm{E}\left[I(\alpha_i = k) | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}\right] I(S_i = s)}$$

and

$$w_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S} \mathrm{E}\left[I(\alpha_i = k) | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}\right] I(S_i = s);$$

$(\boldsymbol{\psi}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)})$ solves the partial likelihood score equation from the full data using one-step Newton-Raphson iteration,

$$\sum_{i=1}^{n}\sum_{s=1}^{S}\Delta_i \left\{ \begin{pmatrix} E[(\widetilde{\boldsymbol{Z}}_i^T(V_i) \circ \boldsymbol{b}_i) | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}] \\ \boldsymbol{Z}_i \end{pmatrix} - \frac{\sum_{l:V_l \geq V_i} \begin{pmatrix} E[(\widetilde{\boldsymbol{Z}}_l^T(V_i) \circ \boldsymbol{b}_l)\exp\{\widetilde{\boldsymbol{Z}}_l(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_l) + \boldsymbol{Z}_l(V_i)\boldsymbol{\gamma}\} | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}] \\ E[\boldsymbol{Z}_l(V_i)\exp\{\widetilde{\boldsymbol{Z}}_l(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_l) + \boldsymbol{Z}_l(V_i)\boldsymbol{\gamma}\} | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}] \end{pmatrix} I(S_l = s)}{\sum_{l:V_l \geq V_i} E[\exp\{\widetilde{\boldsymbol{Z}}_l(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_l) + \boldsymbol{Z}_l(V_i)\boldsymbol{\gamma}\} | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}] I(S_l = s)} \right\} I(S_i = s)$$

$$= \mathbf{0};$$

$\Lambda_s^{(m+1)}$ is obtained as an empirical function with jumps only at the observed failure time,

$$\Lambda_s^{(m+1)}(t) = \sum_{i:V_i \le t} \frac{\Delta_i I(S_i = s)}{\sum_{l:V_l \ge V_i} \mathrm{E}\left[\exp\left\{ \widetilde{\boldsymbol{Z}}_l(V_i)(\boldsymbol{\psi}^{(m+1)} \circ \boldsymbol{b}_l) + \boldsymbol{Z}_l(V_i)\boldsymbol{\gamma}^{(m+1)} \right\} | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)} \right] I(S_l = s)}.$$

The expressions of the conditional expectation and the conditional score equations calculated in the E- and M-steps for continuous longitudinal outcomes following a normal distribution and binary longitudinal outcomes with survival time are given respectively in Appendices A.1 and A.2.

The observed information matrix via Louis (1982) formula is adopted to obtain the variance estimate for $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}(t))$. The variance of $\sqrt{n}\ \widehat{\boldsymbol{\theta}}$ is asymptotically equal to the corresponding sub-matrix of the inverse of the calculated observed information matrix.

4.2.3 EM algorithm – examples

4.2.3.1 Continuous longitudinal data with Normal distribution and survival time

(1) E-step : For continuous longitudinal outcomes following a normal distribution and survival time, we calculate the conditional expectation of $q(\boldsymbol{b}_i, \alpha_i)$ for subject i with $S_i = s$ given the observations and the current estimate $(\boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)})$ for some known function $q(\cdot)$. The conditional expectation denoted by $E\big[q(\boldsymbol{b}_i, \alpha_i) | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}\big]$ can be expressed as the following: Given the current estimate $(\boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)})$,

$$E\big[q(\boldsymbol{b}_i, \alpha_i) | \boldsymbol{\theta}^{(m)}, \Lambda_s^{(m)}\big] = \frac{\sum_{\alpha_i=1}^{K} C_{\alpha_i} \int_{\boldsymbol{z}_{\alpha_i}} q\big(R(\boldsymbol{z}_{\alpha_i})\big) \kappa(\boldsymbol{z}_{\alpha_i}) \exp\{-\boldsymbol{z}_{\alpha_i}^T \boldsymbol{z}_{\alpha_i}\} d\boldsymbol{z}_{\alpha_i}}{\sum_{a=1}^{K} C_a \int_{\boldsymbol{z}_a} \kappa(\boldsymbol{z}_a) \exp\{-\boldsymbol{z}_a^T \boldsymbol{z}_a\} d\boldsymbol{z}_a}, \qquad (4.5)$$

where

$$R(\boldsymbol{z}_{\alpha_i}) = \left(\frac{1}{\sigma_y^2} \sum_{j=1}^{n_i} \widetilde{\boldsymbol{X}}_{ij}^T \widetilde{\boldsymbol{X}}_{ij} + \big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1} \right)^{-1} \left[\sum_{j=1}^{n_i} \frac{1}{\sigma_y^2} \big(y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta}\big) \widetilde{\boldsymbol{X}}_{ij}^T + \Delta_i \big(\widetilde{\boldsymbol{Z}}_i^T(V_i) \circ \boldsymbol{\psi}^{(m)} \big) \right.$$

$$+\big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1}\boldsymbol{\mu}_{\alpha_i}\Big] + \sqrt{2}\Big[\frac{1}{\sigma_y^2}\sum_{j=1}^{n_i}\widetilde{\boldsymbol{X}}_{ij}^T\widetilde{\boldsymbol{X}}_{ij} + \big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1}\Big]^{-\frac{1}{2}}\boldsymbol{z}_{\alpha_i},$$

$$\kappa(\boldsymbol{z}_{\alpha_i}) = \exp\Big\{-\sum_{s=1}^{S} I(S_i = s)\int_0^{V_i} e^{\widetilde{\boldsymbol{Z}}_i(u)\left(\boldsymbol{\psi}^{(m)}\circ R(\boldsymbol{z}_{\alpha_i})\right)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}^{(m)}} d\Lambda_s^{(m)}(u)\Big\},$$

$$C_{\alpha_i} = \exp\Bigg\{\frac{1}{2}\Big[\sum_{j=1}^{n_i}\frac{1}{\sigma_y^2}\big(y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta}\big)\widetilde{\boldsymbol{X}}_{ij}^T + \Delta_i\big(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{\psi}^{(m)}\big) + \big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1}\boldsymbol{\mu}_{\alpha_i}\Big]^T$$
$$\times\left(\frac{1}{\sigma_y^2}\sum_{j=1}^{n_i}\widetilde{\boldsymbol{X}}_{ij}^T\widetilde{\boldsymbol{X}}_{ij} + \big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1}\right)^{-1}\times\Big[\sum_{j=1}^{n_i}\frac{1}{\sigma_y^2}\big(y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta}\big)\widetilde{\boldsymbol{X}}_{ij}^T$$
$$+\ \Delta_i\big(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{\psi}^{(m)}\big) + \big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1}\boldsymbol{\mu}_{\alpha_i}\Big]$$
$$-\frac{1}{2}\boldsymbol{\mu}_{\alpha_i}^T\big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1}\boldsymbol{\mu}_{\alpha_i} + \log w_{\alpha_i}\Bigg\}$$

is a constant,

$\big(\boldsymbol{\Sigma}_b^{(m)}\big)^{\frac{1}{2}}$ is an unique non-negative square root of $\boldsymbol{\Sigma}_b^{(m)}$ (i.e. $\big(\boldsymbol{\Sigma}_b^{(m)}\big)^{\frac{1}{2}}\times\big(\boldsymbol{\Sigma}_b^{(m)}\big)^{\frac{1}{2}} = \boldsymbol{\Sigma}_b^{(m)}$), and $\boldsymbol{z}_{\alpha_i}$ follows a multivariate Gaussian distribution with mean zero.

(2) M-step : Since normal distribution has a dispersion parameter $\boldsymbol{\phi}$ as σ_y^2, we estimate $\boldsymbol{\beta}^{(m+1)}$ and σ_y^2 in longitudinal process. $\boldsymbol{\beta}^{(m+1)}$ is the linear regression coefficients of regressing $\big\{\boldsymbol{Y}_i - \mathrm{E}[\widetilde{\boldsymbol{X}}_i\boldsymbol{b}_i|\boldsymbol{\theta}^{(m)},\boldsymbol{\Lambda}^{(m)}], i = 1,\ldots,n\big\}$ on $\big\{\boldsymbol{X}_i, i = 1,\ldots,n\big\}$, where $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^T,...,\boldsymbol{X}_{in_i}^T)^T$ and $\widetilde{\boldsymbol{X}}_i = (\widetilde{\boldsymbol{X}}_{i1}^T,...,\widetilde{\boldsymbol{X}}_{in_i}^T)^T$.

$$(\sigma_y^2)^{(m+1)} = \frac{\sum_{i=1}^n\big[\boldsymbol{D}_i^T\boldsymbol{D}_i + \mathrm{E}[(\widetilde{\boldsymbol{X}}_i\boldsymbol{b}_i)^2|\boldsymbol{\theta}^{(m)},\boldsymbol{\Lambda}^{(m)}] - \big(\mathrm{E}[\widetilde{\boldsymbol{X}}_i\boldsymbol{b}_i|\boldsymbol{\theta}^{(m)},\boldsymbol{\Lambda}^{(m)}]\big)^2\big]}{\sum_{i=1}^n n_i},$$

where $\boldsymbol{D}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(m+1)} - \mathrm{E}[\widetilde{\boldsymbol{X}}_i\boldsymbol{b}_i|\boldsymbol{\theta}^{(m)},\boldsymbol{\Lambda}^{(m)}]$. $\boldsymbol{\Sigma}_b^{(m+1)}$, $\boldsymbol{\mu}^{(m+1)}$, $\boldsymbol{w}^{(m+1)}$, $(\boldsymbol{\psi}^{(m+1)},\boldsymbol{\gamma}^{(m+1)})$, and $\Lambda_s^{(m+1)}$ have the same expressions as in Section 4.2.2.

4.2.3.2 Binary longitudinal data and survival time

(1) E-step : For binary longitudinal outcomes and survival time, given the current estimate $(\boldsymbol{\theta}^{(m)},\Lambda_s^{(m)})$, the conditional expectation denoted by $E\big[q(\boldsymbol{b}_i,\alpha_i)|\boldsymbol{\theta}^{(m)},\Lambda_s^{(m)}\big]$

can be expressed as in (4.5), where

$$
\begin{aligned}
R(\boldsymbol{z}_{\alpha_i}) &= \boldsymbol{\Sigma}_b^{(m)}\left[\sum_{j=1}^{n_i} y_{ij}\widetilde{\boldsymbol{X}}_{ij}^T + \Delta_i\big(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{\psi}^{(m)}\big)\right] + \boldsymbol{\mu}_{\alpha_i} + \sqrt{2}\big(\boldsymbol{\Sigma}_b^{(m)}\big)^{\frac{1}{2}}\boldsymbol{z}_{\alpha_i},\\
\kappa(\boldsymbol{z}_{\alpha_i}) &= \exp\left\{-\sum_{j=1}^{n_i}\log\left(1+e^{\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(m)}+\widetilde{\boldsymbol{X}}_{ij}R(\boldsymbol{z}_{\alpha_i})}\right)\right.\\
&\qquad\left.-\sum_{s=1}^{S} I(S_i=s)\int_0^{V_i} e^{\widetilde{\boldsymbol{Z}}_i(u)\left(\boldsymbol{\psi}^{(m)}\circ R(\boldsymbol{z}_{\alpha_i})\right)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}^{(m)}} d\Lambda_s^{(m)}(u)\right\}, \text{ and}\\
C_{\alpha_i} &= \exp\left\{\frac{1}{2}\left[\boldsymbol{\Sigma}_b^{(m)}\left(\sum_{j=1}^{n_i} y_{ij}\widetilde{\boldsymbol{X}}_{ij}^T + \Delta_i\big(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{\psi}^{(m)}\big)\right)+\boldsymbol{\mu}_{\alpha_i}\right]^T\times\big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1}\right.\\
&\qquad\times\left[\boldsymbol{\Sigma}_b^{(m)}\left(\sum_{j=1}^{n_i} y_{ij}\widetilde{\boldsymbol{X}}_{ij}^T + \Delta_i\big(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{\psi}^{(m)}\big)\right)+\boldsymbol{\mu}_{\alpha_i}\right]\\
&\qquad\left.-\frac{1}{2}\boldsymbol{\mu}_{\alpha_i}^T\big(\boldsymbol{\Sigma}_b^{(m)}\big)^{-1}\boldsymbol{\mu}_{\alpha_i}+\log w_{\alpha_i}\right\}
\end{aligned}
$$

is a constant.

(2) M-step : Since the parameter $\boldsymbol{\phi}$ is set to 1 for logistic distribution, we estimate only $\boldsymbol{\beta}$ in the longitudinal process. $\boldsymbol{\beta}^{(m+1)}$ solves the conditional expectation of complete data log-likelihood score equation, using one-step Newton-Raphson iteration,

$$
\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij}-\sum_{s=1}^{S}\mathrm{E}\left[\frac{\exp\{\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(m+1)}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\}}{1+\exp\{\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(m+1)}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\}}\middle|\boldsymbol{\theta}^{(m)},\Lambda_s^{(m)}\right]I(S_i=s)\right)\boldsymbol{X}_{ij}^T \;=\; \mathbf{0}.
$$

$\boldsymbol{\Sigma}_b^{(m+1)}, \boldsymbol{\mu}^{(m+1)}, \boldsymbol{w}^{(m+1)}, \boldsymbol{\psi}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}$, and $\Lambda_s^{(m+1)}$ have the same expressions as in Section 4.2.2.

4.3 Asymptotic Properties

To study the asymptotic properties of the proposed estimator $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}(t))$ with $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\phi}}^T, \mathrm{Vec}(\widehat{\boldsymbol{\Sigma}}_b)^T, \boldsymbol{\mu}^T, \boldsymbol{w}^T, \widehat{\boldsymbol{\psi}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ and $\widehat{\boldsymbol{\Lambda}}(t) = (\widehat{\Lambda}_1(t), \ldots, \widehat{\Lambda}_S(t))^T$, we assume the following conditions.

(A1) The true parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\phi}_0^T, \mathrm{Vec}(\boldsymbol{\Sigma}_{b0})^T, \boldsymbol{\mu}^T, \boldsymbol{w}^T, \boldsymbol{\psi}_0^T, \boldsymbol{\gamma}_0^T)^T$ belongs to a known

compact set Θ which lies in the interior of the domain for $\boldsymbol{\theta}$.

(A2) The distribution of random effects $\boldsymbol{b}_i^*$ is a mixture of a finite number of d_b-dimensional multivariate normal distributions with means $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_K^T)^T$ and a common covariance matrix $\boldsymbol{\Sigma}_b$. i.e. $\boldsymbol{b}_i^* \sim \sum_{k=1}^K w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_b)$, where K is the number of mixture components.

(A3) The true baseline hazard rate function $\boldsymbol{\lambda}_0(t) = (\lambda_{10}(t), \ldots, \lambda_{S0}(t))$ is continuous and positive in $[0, \tau]$, where τ is the time of study end.

(A4) For the censoring time C, $P(C \geq \tau | \boldsymbol{Z}, \widetilde{\boldsymbol{Z}}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}) = P(C = \tau | \boldsymbol{Z}, \widetilde{\boldsymbol{Z}}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}) > 0$.

(A5) For the number of observed longitudinal measurements per subject n_N, $P(n_N > d_b | \boldsymbol{X}, \widetilde{\boldsymbol{X}}) > 0$ with probability one, and $P(n_N \leq n_0) = 1$ for some integer n_0.

(A6) Both $\boldsymbol{X}^T\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}$ are full rank with positive probability. Moreover, if there exist constant vectors $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ such that, with positive probability, for any t, $\boldsymbol{Z}(t)\boldsymbol{c}_1 = \alpha_0(t)$ and $\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{c}_2 = 0$ for a deterministic function $\alpha_0(t)$, then $\boldsymbol{c}_1 = 0$, $\boldsymbol{c}_2 = 0$, and $\alpha_0(t) = 0$.

Assumption (A4) means that, by the end of the study, some proportion of the subjects will still be alive and censored at the study end time τ, and thus the maximum right censoring time is equal to τ. Assumption (A5) implies that some proportion of the subjects have at least d_b longitudinal observations, and there exists an integer n_0 such that all subjects have a finite number of longitudinal observations which are not larger than n_0. Consistency and asymptotic distribution of the proposed estimator are summarized in the following two theorems. We will present outlines of the proofs here. The detailed technical proofs are given in Section 4.4.

Theorem 4.1. *Under the assumptions (A1)~(A6), as $n \to \infty$, the maximum likelihood estimator $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}(t))$ is consistent under the product norm of the Euclidean distance and*

the supreme norm on $[0,\tau]$. *That is,* $\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|+\sup_{t\in[0,\tau]}\|\widehat{\boldsymbol{\Lambda}}(t)-\boldsymbol{\Lambda}_0(t)\|\longrightarrow 0$, *a.s., where* $\|\widehat{\boldsymbol{\Lambda}}(t)-\boldsymbol{\Lambda}_0(t)\|=\sum_{s=1}^{S}|\widehat{\Lambda}_s(t)-\Lambda_{s0}(t)|$.

Consistency in Theorem 4.1 can be proved by verifying the following three steps: First, we show that the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}})$ exists. Second, we show that, with probability one, $\widehat{\Lambda}_s(\tau)$, $s=1,\ldots,S$, are bounded as $n\to\infty$. Third, given that the second step is true, by Helly's selection theorem (van der Vaart, 1998), we can choose a subsequence of $\widehat{\Lambda}_s(t)$ such that $\widehat{\Lambda}_s(t)$ weakly converges to some right-continuous monotone function $\Lambda_s^*(t)$ with probability one. Also, for any sub-sequence, we can find a further sub-sequence, still denoted as $\widehat{\boldsymbol{\theta}}$, such that $\widehat{\boldsymbol{\theta}}\to\boldsymbol{\theta}^*$. Using empirical process formulation and relevant Donsker properties with parameter identifiability, we can show that $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$ and $\Lambda_s^*=\boldsymbol{\Lambda}_{s0}$, $s=1,\ldots,S$. Once the three steps are completed, we can conclude that, with probability one, $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\Lambda}}_s(t)$ converges to $\boldsymbol{\Lambda}_{s0}(t)$ in $[0,\tau]$, $s=1,\ldots,S$. Moreover, since $\boldsymbol{\Lambda}_{s0}(t)$ is right-continuous in $[0,\tau]$, the latter can be strengthened to uniform convergence; that is, $\sup_{t\in[0,\tau]}\|\widehat{\boldsymbol{\Lambda}}(t)-\boldsymbol{\Lambda}_0(t)\|\to 0$ almost surely.

Theorem 4.2. *Under the assumptions (A1)~(A6), as* $n\to\infty$, $\sqrt{n}\big((\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T,(\widehat{\boldsymbol{\Lambda}}(t)-\boldsymbol{\Lambda}_0(t))^T\big)^T$ *weakly converges to a Gaussian random element in* $R^{d_\theta}\times\ell^\infty[0,\tau]\times\cdots\times\ell^\infty[0,\tau]$, *and the estimator* $\widehat{\boldsymbol{\theta}}$ *is asymptotically efficient, where* d_θ *is the dimension of* $\boldsymbol{\theta}$ *and* $\ell^\infty[0,\tau]$ *is the normed space containing all the bounded functions in* $[0,\tau]$.

Once consistency holds, the conditions of Theorem 3.3.1 in van der Vaart and Wellner (1996), which implies the asymptotic normality in Theorem 4.2, are verified via the tools of empirical processes. These conditions are restated in Theorem 4 of Parner (1998). The smooth conditions in Theorem 4 of Parner (1998) can be verified using the regularity of the log-likelihood function in terms of model parameters and the Donsker properties of the score operators. In particular, in the invertibility condition of the

information operator in Theorem 4 of Parner (1998), the verification of the one-to-one property of the information operator is specific to our proposed models and requires non-trivial work. Therefore, by Theorem 3.3.1 of van der Vaart and Wellner (1996), $\sqrt{n}\big((\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T, (\widehat{\boldsymbol{\Lambda}}(t)-\boldsymbol{\Lambda}_0(t))^T\big)^T$ weakly converges to a Gaussian process, and by Proposition 3.3.1 in Bickel *et al.* (1993), $\widehat{\boldsymbol{\theta}}$ is an efficient estimator for $\boldsymbol{\theta}_0$.

4.4 Technical Details – Proofs for Asymptotic Properties

In this section, we present the detailed technical proofs for the asymptotic properties of the proposed estimator $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}(t))$ with $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\phi}}^T, \mathrm{Vec}(\widehat{\boldsymbol{\Sigma}}_b)^T, \widehat{\boldsymbol{\mu}}^T, \widehat{\boldsymbol{w}}^T, \widehat{\boldsymbol{\psi}}^T, \widehat{\boldsymbol{\gamma}}^T)^T$ and $\widehat{\boldsymbol{\Lambda}}(t) = (\widehat{\Lambda}_1(t), \ldots, \widehat{\Lambda}_S(t))^T$. Meanwhile, the supplementary proofs needed to prove the asymptotic properties are provided in Section 4.4.3. We use $\boldsymbol{b}$ to denote random effects instead of $\boldsymbol{b}^*$ for convenience in all proofs. From the following full likelihood function of observed data $(\boldsymbol{Y}, \boldsymbol{V})$ for $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$,

$$
\begin{aligned}
&L_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) \\
&= \sum_{\boldsymbol{\alpha}} \int_{\boldsymbol{b}} L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}, \boldsymbol{\alpha}) d\boldsymbol{b} \\
&= \prod_{i=1}^{n} \Bigg(\sum_{\alpha_i=1}^{K} \int_{\boldsymbol{b}_1} \cdots \int_{\boldsymbol{b}_K} L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}_i, V_i, \boldsymbol{b}_{\alpha_i}, \alpha_i) d\boldsymbol{b}_1 \cdots d\boldsymbol{b}_K \Bigg) \\
&= \prod_{i=1}^{n} \Bigg(\sum_{\alpha_i=1}^{K} \int_{\boldsymbol{b}_{\alpha_i}} L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}_i, V_i, \boldsymbol{b}_{\alpha_i}, \alpha_i) d\boldsymbol{b}_{\alpha_i} \Bigg) \\
&= \prod_{i=1}^{n} \Bigg(\sum_{\alpha_i=1}^{K} \prod_{k=1}^{K} \Bigg[\int_{\boldsymbol{b}_k} \Bigg(\exp\Bigg\{ \sum_{j=1}^{n_i} \Bigg[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_{ik}) - B(\boldsymbol{\beta}; \boldsymbol{b}_{ik})}{A(D_i(t_j;\phi))} + C(Y_{ij}; D_i(t_j;\phi)) \Bigg] \Bigg\} \\
&\qquad \times \Bigg(\prod_{s=1}^{S} \Bigg[\lambda_s(V_i)^{\Delta_i} \exp\Bigg\{ \Delta_i \big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_{ik}) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma} \big] \\
&\qquad\qquad - \int_0^{V_i} \exp\big\{ \widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_{ik}) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma} \big\} d\Lambda_s(u) \Bigg\} \Bigg]^{I(S_i=s)} \Bigg) \\
&\qquad \times (2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\Bigg\{ -\frac{1}{2} (\boldsymbol{b}_{ik} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_b^{-1} (\boldsymbol{b}_{ik} - \boldsymbol{\mu}_k) \Bigg\} \times w_k \Bigg) d\boldsymbol{b}_k \Bigg]^{I(\alpha_i=k)} \Bigg),
\end{aligned}
$$

we have the observed log-likelihood function

$$\sum_{i=1}^{n}\log\Bigg(\sum_{\alpha_i=1}^{K}\prod_{k=1}^{K}\Bigg[\int_{\boldsymbol{b}_k}\Bigg(\exp\Bigg\{\sum_{j=1}^{n_i}\Bigg[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_{ik})-B(\boldsymbol{\beta};\boldsymbol{b}_{ik})}{A(D_i(t_j;\phi))}+C(Y_{ij};D_i(t_j;\phi))\Bigg]\Bigg\}$$
$$\times\Bigg(\prod_{s=1}^{S}\Bigg[\lambda_s(V_i)^{\Delta_i}\exp\Bigg\{\Delta_i\big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi}\circ\boldsymbol{b}_{ik})+\boldsymbol{Z}_i(V_i)\boldsymbol{\gamma}\big]$$
$$-\int_0^{V_i}\exp\big\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_{ik})+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}\big\}d\Lambda_s(u)\Bigg\}\Bigg]^{I(S_i=s)}\Bigg)$$
$$\times(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\Bigg\{-\frac{1}{2}(\boldsymbol{b}_{ik}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_b^{-1}(\boldsymbol{b}_{ik}-\boldsymbol{\mu}_k)\Bigg\}\times w_k\Bigg)d\boldsymbol{b}_k\Bigg]^{I(\alpha_i=k)}\Bigg).$$

Then, we obtain the following modified object function by replacing $\lambda_s(V_i)$ with $\Lambda_s\{V_i\}$ in the above expression where $\Lambda_s\{V_i\}$ is the jump size of $\Lambda_s(t)$ at the observed time V_i with $\Delta_i = 1$,

$$\begin{aligned}
&l_n(\boldsymbol{\theta},\boldsymbol{\Lambda})\\
&=\sum_{i=1}^{n}\log\Bigg(\sum_{\alpha_i=1}^{K}\prod_{k=1}^{K}\Bigg[\int_{\boldsymbol{b}_k}\Bigg(\exp\Bigg\{\sum_{j=1}^{n_i}\Bigg[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_{ik})-B(\boldsymbol{\beta};\boldsymbol{b}_{ik})}{A(D_i(t_j;\phi))}+C(Y_{ij};D_i(t_j;\phi))\Bigg]\Bigg\}\\
&\qquad\times\Bigg(\prod_{s=1}^{S}\Bigg[\Lambda_s\{V_i\}^{\Delta_i}\exp\Bigg\{\Delta_i\big[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi}\circ\boldsymbol{b}_{ik})+\boldsymbol{Z}_i(V_i)\boldsymbol{\gamma}\big]\\
&\qquad\quad-\int_0^{V_i}\exp\big\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_{ik})+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}\big\}d\Lambda_s(u)\Bigg\}\Bigg]^{I(S_i=s)}\Bigg)\\
&\qquad\times(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\Bigg\{-\frac{1}{2}(\boldsymbol{b}_{ik}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_b^{-1}(\boldsymbol{b}_{ik}-\boldsymbol{\mu}_k)\Bigg\}\times w_k\Bigg)d\boldsymbol{b}_k\Bigg]^{I(\alpha_i=k)}\Bigg),
\end{aligned}\tag{4.6}$$

and $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}})$ maximizes $l_n(\boldsymbol{\theta},\boldsymbol{\Lambda})$ over the space $\{(\boldsymbol{\theta},\boldsymbol{\Lambda}) : \boldsymbol{\theta}\in\Theta,\boldsymbol{\Lambda}\in\mathbb{W}_n\times\mathbb{W}_n\cdots\times\mathbb{W}_n\}$, where $\mathbb{W}_n$ consists of all the right-continuous step functions only; that is, $\boldsymbol{\Lambda} = (\Lambda_1,\ldots,\Lambda_S)^T, s = 1,\ldots,S,\Lambda_s\in\mathbb{W}_n$. For the proofs of both Theorem 4.1 and Theorem 4.2, the modified object function is used in the place of the observed log-likelihood function.

4.4.1 Proof of consistency

Consistency can be proved by verifying the following three steps: First, we show the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}})$ exists. Second, we show that, with probability one, $\widehat{\Lambda}_s(\tau)$, $s = 1, \ldots, S$, are bounded as $n \to \infty$. Third, if the second step is true, by Helly's selection theorem (p9 of van der Vaart (1998), we can choose a subsequence of $\widehat{\Lambda}_s$ such that $\widehat{\Lambda}_s$ weakly converges to some right-continuous monotone function Λ_s^* with probability one; that is, the measure given by $\mu_s([0,t]) = \widehat{\Lambda}_s(t)$ for $t \in [0,\tau]$ weakly converges to the measure given by $\mu_s^*([0,t]) = \Lambda_s^*(t)$. By choosing a sub-sequence, we can further assume $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$. Thus, in this third step, we show $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda_s^* = \Lambda_{s0}$, $s = 1, \ldots, S$.

Once the three steps are completed, we can conclude that, with probability one, $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\Lambda}}_s$ converges to $\boldsymbol{\Lambda}_{s0}$ in $[0,\tau]$, $s = 1, \ldots, S$. However, since $\boldsymbol{\Lambda}_{s0}$ is continuous in $[0,\tau]$, the latter can be strengthened to uniform convergence; that is, $\sup_{t\in[0,\tau]} \|\widehat{\boldsymbol{\Lambda}}(t) - \boldsymbol{\Lambda}_0(t)\| \to 0$ almost surely. Then, the proof of Theorem 4.1 will be done.

In the first step, we will show the existence of the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}})$. Since $\boldsymbol{\theta}$ belongs to a compact set Θ by the assumption (A1), it is sufficient to show that $\Lambda_s\{V_i\}$, the jump size of Λ_s at V_i for which $\Delta_i = 1$, is finite. Since, for each subject i with $\Delta_i = 1$,

$$\begin{aligned}
&\Lambda_s\{V_i\} \exp\left\{ - \int_0^{V_i} \exp\left\{\widetilde{\boldsymbol{Z}}_i(t)(\boldsymbol{\psi} \circ \boldsymbol{b}) + \boldsymbol{Z}_i(t)\boldsymbol{\gamma}\right\} d\Lambda_s(t) \right\} \\
&\qquad \leq \exp\left\{ -2\big(\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma}\big)\right\} \big(\Lambda_s\{V_i\}\big)^{-1},
\end{aligned}$$

we have that, from (4.6),

$$\begin{aligned}
&l_n(\boldsymbol{\theta}, \boldsymbol{\Lambda}) \\
&\leq \sum_{i=1}^{n} \log\left(\sum_{\alpha_i=1}^{K} \prod_{k=1}^{K} \left[\int_{\boldsymbol{b}_k} \left(\exp\left\{ \sum_{j=1}^{n_i} \left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_k) - B(\boldsymbol{\beta}; \boldsymbol{b}_k)}{A(D_i(t_j;\phi))} + C(Y_{ij}; D_i(t_j;\phi)) \right] \right\} \right. \right. \right.
\end{aligned}$$

$$\times \left(\prod_{s=1}^{S} \left[\left(\Lambda_s\{V_i\} \right)^{-\Delta_i} \exp\left\{ -\Delta_i \left[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_k) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma} \right] \right\} \right]^{I(S_i=s)} \right)$$

$$\times (2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp\left\{ -\frac{1}{2}(\boldsymbol{b}_k - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_b^{-1} (\boldsymbol{b}_k - \boldsymbol{\mu}_k) \right\} \times w_k \Big) d\boldsymbol{b}_k \Bigg]^{I(\alpha_i=k)} \Bigg).$$

Thus, if $\Lambda_s\{V_i\} \to \infty$ for some i with $\Delta_i = 1$, then $l_n(\boldsymbol{\theta}, \boldsymbol{\Lambda}) \to -\infty$, which is contradictory to that $l_n(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ is bounded. Therefore, we conclude that $\Lambda_s\{\cdot\}$, the jump size of Λ_s for stratum s, must be finite. By the conclusion and the assumption (A1), the maximum likelihood estimate $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}})$ exists.

In the second step, we will show that $\widehat{\Lambda}_s(\tau)$ is bounded as n goes to infinity with probability one. We define $\widehat{\zeta}_s = \log \widehat{\Lambda}_s(\tau)$ and rescale $\widehat{\Lambda}_s$ by the factor $e^{\widehat{\zeta}_s}$. Then, we let $\widetilde{\Lambda}_s$ denote the rescaled function; that is, $\widetilde{\Lambda}_s(t) = \widehat{\Lambda}_s(t)/\widehat{\Lambda}_s(\tau) = \widehat{\Lambda}_s(t)e^{-\widehat{\zeta}_s}$. thus, $\widetilde{\Lambda}_s(\tau) = 1$. To prove this second step, it is sufficient to show $\widehat{\zeta}_s$ is bounded. After some algebra in (4.6), we obtain that, for any $\boldsymbol{\Lambda} \in \mathbb{W} \times \mathbb{W} \cdots \times \mathbb{W}$,

$$\begin{aligned}
& n^{-1} l_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\Lambda}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \Bigg[\sum_{j=1}^{n_i} \left(\frac{Y_{ij} \boldsymbol{X}_{ij} \widehat{\boldsymbol{\beta}}}{A(D_i(t_j; \widehat{\phi}))} + C(Y_{ij}; D_i(t_j; \widehat{\phi})) \right) + \sum_{s=1}^{S} \Delta_i I(S_i = s) \log \Lambda_s\{V_i\} + \Delta_i \boldsymbol{Z}_i(V_i) \widehat{\boldsymbol{\gamma}} \\
&\quad - \frac{1}{2} \log\left\{ (2\pi)^{d_b} |\widehat{\boldsymbol{\Sigma}}_b| \right\} - \frac{1}{2} \log |\widehat{\boldsymbol{\Sigma}}_b| + \log \Bigg[\sum_{\alpha=1}^{K} \Bigg[\widehat{w}_\alpha \exp\left\{ \frac{1}{2} \boldsymbol{M}_{i\alpha}^T \boldsymbol{M}_{i\alpha} - \frac{1}{2} \widehat{\boldsymbol{\mu}}_\alpha^T \widehat{\boldsymbol{\Sigma}}_b^{-1} \widehat{\boldsymbol{\mu}}_\alpha \right\} \\
&\qquad \times \int_{\boldsymbol{b}_{\alpha 0}} \Bigg[\exp\Bigg\{ -\frac{1}{2} \boldsymbol{b}_{\alpha 0}^T \boldsymbol{b}_{\alpha 0} - \sum_{j=1}^{n_i} \frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_{\alpha 0})}{A(D_i(t_j; \widehat{\phi}))} \\
&\qquad\quad - \sum_{s=1}^{S} I(S_i = s) \int_0^{V_i} e^{Q_{1i\alpha}(t, \boldsymbol{b}_{\alpha 0}, \widehat{\boldsymbol{\theta}})} d\Lambda_s(t) \Bigg\} \Bigg] d\boldsymbol{b}_{\alpha 0} \Bigg] \Bigg] \Bigg],
\end{aligned}$$

where
$$\boldsymbol{M}_{i\alpha} = \left[\left(\sum_{j=1}^{n_i} \frac{Y_{ij} \widetilde{\boldsymbol{X}}_{ij}}{A(D_i(t_j; \widehat{\phi}))} + \Delta_i \left(\widetilde{\boldsymbol{Z}}_i(V_i) \circ \widehat{\boldsymbol{\psi}}^T \right) \right) \widehat{\boldsymbol{\Sigma}}_b^{1/2} + \widehat{\boldsymbol{\mu}}_\alpha^T \widehat{\boldsymbol{\Sigma}}_b^{-1/2} \right]^T,$$
$$\boldsymbol{b}_{\alpha 0} = \boldsymbol{\Sigma}_b^{-1/2} \boldsymbol{b}_\alpha - \boldsymbol{M}_{i\alpha},$$
and
$$Q_{1i\alpha}(t, \boldsymbol{b}_{\alpha 0}, \widehat{\boldsymbol{\theta}}) = \left(\widetilde{\boldsymbol{Z}}_i(t) \circ \widehat{\psi}^T \right) \widehat{\boldsymbol{\Sigma}}_b^{1/2} \boldsymbol{b}_{\alpha 0} + \boldsymbol{Z}_i(t) \widehat{\boldsymbol{\gamma}} + \left(\widetilde{\boldsymbol{Z}}_i(t) \circ \widehat{\psi}^T \right) \widehat{\boldsymbol{\Sigma}}_b^{1/2} \boldsymbol{M}_{i\alpha}.$$

Thus, since $0 \le n^{-1}l_n(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}) - n^{-1}l_n(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\Lambda}})$ where $\widehat{\boldsymbol{\Lambda}} = e^{\widehat{\boldsymbol{\xi}}} \circ \widetilde{\boldsymbol{\Lambda}}$, it follows that

$$
\begin{aligned}
0 \;\le\; & \frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S} I(S_i = s)\Delta_i\Big(\log e^{\widehat{\zeta}_s}\widetilde{\Lambda}_s\{V_i\} - \log \widetilde{\Lambda}_s\{V_i\}\Big) \\
& + \frac{1}{n}\sum_{i=1}^{n}\log\Bigg[\sum_{\alpha=1}^{K}\Bigg[\widehat{w}_\alpha \exp\Big\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\Big\} \\
& \qquad \times \int_{\boldsymbol{b}_{\alpha0}} \exp\Bigg\{ -\frac{1}{2}\boldsymbol{b}_{\alpha0}^T\boldsymbol{b}_{\alpha0} - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_{\alpha0})}{A(D_i(t_j;\widehat{\phi}))} \\
& \qquad\qquad - \sum_{s=1}^{S} I(S_i = s)e^{\widehat{\zeta}_s}\int_0^{V_i} e^{Q_{1i\alpha}(t, \boldsymbol{b}_{\alpha0}, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t)\Bigg\}\boldsymbol{b}_{\alpha0}\Bigg]\Bigg] \\
& - \frac{1}{n}\sum_{i=1}^{n}\log\Bigg[\sum_{\alpha=1}^{K}\Bigg[\widehat{w}_\alpha \exp\Big\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\Big\} \\
& \qquad \times \int_{\boldsymbol{b}_{\alpha0}} \exp\Bigg\{ -\frac{1}{2}\boldsymbol{b}_{\alpha0}^T\boldsymbol{b}_{\alpha0} - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_{\alpha0})}{A(D_i(t_j;\widehat{\phi}))} \\
& \qquad\qquad - \sum_{s=1}^{S} I(S_i = s)\int_0^{V_i} e^{Q_{1i\alpha}(t, \boldsymbol{b}_{\alpha0}, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t)\Bigg\}d\boldsymbol{b}_{\alpha0}\Bigg]\Bigg]. \qquad (4.7)
\end{aligned}
$$

According to the assumption (A3), there exist some positive constants $C1$, $C2$ and $C3$ such that $|Q_{1i\alpha}(t, \boldsymbol{b}_{\alpha0}, \widehat{\boldsymbol{\theta}})| \le C_1\|\boldsymbol{b}_{\alpha0}\| + C_2\|\boldsymbol{Y}_i\| + C_3$. By denoting $\boldsymbol{b}_{\alpha0}$ as a vector of variables following a standard multivariate normal distribution, from concavity of the logarithm function, in the third term of (4.7),

$$
\begin{aligned}
& \log\Bigg[\sum_{\alpha=1}^{K}\Bigg[\widehat{w}_\alpha \exp\Big\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\Big\} \\
& \qquad \times \int_{\boldsymbol{b}_{\alpha0}} \exp\Bigg\{ -\frac{1}{2}\boldsymbol{b}_{\alpha0}^T\boldsymbol{b}_{\alpha0} - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_{\alpha0})}{A(D_i(t_j;\widehat{\phi}))} - \sum_{s=1}^{S} I(S_i = s)\int_0^{V_i} e^{Q_{1i\alpha}(t, \boldsymbol{b}_{\alpha0}, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t)\Bigg\}d\boldsymbol{b}_{\alpha0}\Bigg]\Bigg] \\
& = \frac{d_b}{2}\log(2\pi) + \log\Bigg[\mathrm{E}_\alpha\Bigg[\exp\Big\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\Big\} \\
& \qquad \times \mathrm{E}_{\boldsymbol{b}_{\alpha0}|\alpha}\Bigg[\exp\Bigg\{ -\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_{\alpha0})}{A(D_i(t_j;\widehat{\phi}))} - \sum_{s=1}^{S} I(S_i = s)\int_0^{V_i} e^{Q_{1i\alpha}(t, \boldsymbol{b}_{\alpha0}, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t)\Bigg\}\Bigg]\Bigg]\Bigg] \\
& = \frac{d_b}{2}\log(2\pi) + \log\Bigg[\mathrm{E}_{\alpha,\boldsymbol{b}_0}\Bigg[\exp\Big\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\Big\} \\
& \qquad \times \exp\Bigg\{ -\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}}; \boldsymbol{b}_{\alpha0})}{A(D_i(t_j;\widehat{\phi}))} - \sum_{s=1}^{S} I(S_i = s)\int_0^{V_i} e^{Q_{1i\alpha}(t, \boldsymbol{b}_{\alpha0}, \widehat{\boldsymbol{\theta}})} d\widetilde{\Lambda}_s(t)\Bigg\}\Bigg]\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{d_b}{2}\log(2\pi) + \log\Bigg[\mathrm{E}_{\alpha,\boldsymbol{b}_0}\Bigg[\exp\Bigg\{ \frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha \\
&\qquad\qquad - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))} - e^{C_1\|\boldsymbol{b}_{\alpha 0}\|+C_2\|\boldsymbol{Y}_i\|+C_3}\Bigg\}\Bigg]\Bigg] \\
&\geq \frac{d_b}{2}\log(2\pi) + \log\Bigg[\mathrm{E}_{\alpha,\boldsymbol{b}_0}\Bigg[\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))} - e^{C_1\|\boldsymbol{b}_{\alpha 0}\|+C_2\|\boldsymbol{Y}_i\|+C_3}\Bigg]\Bigg] \\
&= -e^{C_2\|\boldsymbol{Y}_i\|+C_4} - C_5,
\end{aligned}
$$

where C_4 and C_5 are positive constants. Then, since it is easily verified that $\mathrm{E}_{\alpha,\boldsymbol{b}_0}\Bigg[\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))} - e^{C_1\|\boldsymbol{b}_{\alpha 0}\|+C_2\|\boldsymbol{Y}_i\|+C_3}\Bigg] < \infty$, by the strong law of large numbers and the assumption (A5), the third term of (4.7)

$$
\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\log\Bigg[\sum_{\alpha=1}^{K}\Bigg[\widehat{w}_\alpha\exp\Bigg\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\Bigg\} \\
&\quad\times\int_{\boldsymbol{b}_{\alpha 0}}\exp\Bigg\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0} - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))} - \sum_{s=1}^{S}I(S_i=s)\int_0^{V_i}e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\Bigg\}d\boldsymbol{b}_{\alpha 0}\Bigg]\Bigg] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(e^{C_2\|\boldsymbol{Y}_i\|+C_4} + C_5) \;\triangleq\; C_6
\end{aligned}
$$

can be bounded by some constant C_6 from above. Then (4.7) becomes

$$
\begin{aligned}
0 \leq\; &\frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\Delta_i I(S_i=s)\widehat{\zeta}_s \\
&+\frac{1}{n}\sum_{i=1}^{n}\log\Bigg[\sum_{\alpha=1}^{K}\Bigg[\widehat{w}_\alpha\exp\Bigg\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\Bigg\} \\
&\quad\times\int_{\boldsymbol{b}_{\alpha 0}}\exp\Bigg\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0} - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))} - \sum_{s=1}^{S}I(S_i=s)e^{\widehat{\zeta}_s}\int_0^{V_i}e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\Bigg\}\boldsymbol{b}_{\alpha 0}\Bigg]\Bigg] \\
&+C_6 \\
\leq\; &\frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\Delta_i I(S_i=s)\widehat{\zeta}_s \\
&+\frac{1}{n}\sum_{i=1}^{n}I(V_i=\tau)\log\Bigg[\sum_{\alpha=1}^{K}\Bigg[\widehat{w}_\alpha\exp\Bigg\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\Bigg\} \\
&\quad\times\int_{\boldsymbol{b}_{\alpha 0}}\exp\Bigg\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0} - \sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))} - \sum_{s=1}^{S}I(S_i=s)e^{\widehat{\zeta}_s}\int_0^{\tau}e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\Bigg\}\boldsymbol{b}_{\alpha 0}\Bigg]\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{n}\sum_{i=1}^{n} I(V_i \neq \tau)\log\left[\sum_{\alpha=1}^{K}\left[\widehat{w}_\alpha \exp\left\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha}-\frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\right\}\right.\right.\\
&\qquad\left.\left.\times\int_{\boldsymbol{b}_{\alpha 0}}\exp\left\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0}-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}\right\}\boldsymbol{b}_{\alpha 0}\right]\right]\\
&+C_6\\
\leq\ &\frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{S}\Delta_i I(S_i=s)\widehat{\zeta}_s\\
&+\frac{1}{n}\sum_{i=1}^{n} I(V_i=\tau)\log\left[\sum_{\alpha=1}^{K}\left[\widehat{w}_\alpha \exp\left\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha}-\frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\right\}\right.\right.\\
&\quad\left.\left.\times\int_{\boldsymbol{b}_{\alpha 0}}\exp\left\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0}-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}-\sum_{s=1}^{S}I(S_i=s)e^{\widehat{\zeta}_s}\int_0^\tau e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\right\}\boldsymbol{b}_{\alpha 0}\right]\right]\\
&+C_7, \qquad (4.8)
\end{aligned}
$$

where C_7 is a constant. On the other hand, since, for any $\Gamma \geq 0$ and $x > 0$, $\Gamma\log\left(1+x/\Gamma\right) \leq \Gamma x/\Gamma = x$, we have that $e^{-x} \leq \left(1+x/\Gamma\right)^{-\Gamma}$. Therefore, in the second term of (4.8),

$$
\begin{aligned}
&\exp\left\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0}-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}-\sum_{s=1}^{S}I(S_i=s)e^{\widehat{\zeta}_s}\int_0^\tau e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\right\}\\
&\leq \exp\left\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0}-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}\right\}\times\left\{1+\frac{\sum_{s=1}^{S}I(S_i=s)e^{\widehat{\zeta}_s}\int_0^\tau e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)}{\Gamma}\right\}^{-\Gamma}\\
&\leq \Gamma^\Gamma\times\exp\left\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0}-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}\right\}\times\left\{\sum_{s=1}^{S}I(S_i=s)e^{\widehat{\zeta}_s}\int_0^\tau e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\right\}^{-\Gamma}.
\end{aligned}
\qquad (4.9)
$$

Since $Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}}) \geq -C_1\|\boldsymbol{b}_{\alpha 0}\| - C_2\|\boldsymbol{Y}_i\| - C_3$,

$$
\begin{aligned}
\int_0^\tau e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t) &\geq \int_0^\tau e^{-C_1\|\boldsymbol{b}_{\alpha 0}\|-C_2\|\boldsymbol{Y}_i\|-C_3}d\widetilde{\Lambda}_s(t)\\
&= e^{-C_1\|\boldsymbol{b}_{\alpha 0}\|-C_2\|\boldsymbol{Y}_i\|-C_3}\times\left\{\widetilde{\Lambda}_s(\tau)-\widetilde{\Lambda}_s(0)\right\}\\
&= e^{-C_1\|\boldsymbol{b}_{\alpha 0}\|-C_2\|\boldsymbol{Y}_i\|-C_3}.
\end{aligned}
$$

Thus, in (4.9),

$$\left\{\sum_{s=1}^{S} I(S_i=s)e^{\widehat{\zeta}_s}\int_0^{\tau} e^{Q_{1i\alpha}(t,\boldsymbol{b}_{\alpha 0},\widehat{\boldsymbol{\theta}})}d\widetilde{\Lambda}_s(t)\right\}^{-\Gamma} \le \left\{\sum_{s=1}^{S} I(S_i=s)e^{\widehat{\zeta}_s}\right\}^{-\Gamma} e^{C_1\Gamma\|\boldsymbol{b}_{\alpha 0}\|+C_2\Gamma\|\boldsymbol{Y}_i\|+C_3\Gamma}$$

and $$(4.9) \le \Gamma^{\Gamma}\exp\left\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0}-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}-\Gamma\log\left(\sum_{s=1}^{S} I(S_i=s)e^{\widehat{\zeta}_s}\right) + C_1\Gamma\|\boldsymbol{b}_{\alpha 0}\|+C_2\Gamma\|\boldsymbol{Y}_i\|+C_3\Gamma\right\}.$$

Therefore, (4.8) gives that

$$\begin{aligned}
0 \le\ & C_7+\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left(\sum_{s=1}^{S} I(S_i=s)\widehat{\zeta}_s\right)\\
&+\frac{1}{n}\sum_{i=1}^{n} I(V_i=\tau)\log\left[\sum_{\alpha=1}^{K}\left[\widehat{w}_\alpha\exp\left\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha}-\frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\right\}\right.\right.\\
&\quad\times\int_{\boldsymbol{b}_{\alpha 0}}\Gamma^{\Gamma}\exp\left\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0}-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}-\Gamma\sum_{s=1}^{S} I(S_i=s)\widehat{\zeta}_s\right.\\
&\qquad\left.\left.+C_1\Gamma\|\boldsymbol{b}_{\alpha 0}\|+C_2\Gamma\|\boldsymbol{Y}_i\|+C_3\Gamma\right\}d\boldsymbol{b}_{\alpha 0}\right]\Bigg]\\
=\ & C_7+\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left(\sum_{s=1}^{S} I(S_i=s)\widehat{\zeta}_s\right)\\
&+\frac{1}{n}\sum_{i=1}^{n} I(V_i=\tau)\log\left[\Gamma^{\Gamma}\exp\left\{-\Gamma\sum_{s=1}^{S} I(S_i=s)\widehat{\zeta}_s\right\}\sum_{\alpha=1}^{K}\left[\widehat{w}_\alpha\exp\left\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha}-\frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha\right\}\right.\right.\\
&\quad\times(2\pi)^{d_b/2}(2\pi)^{-d_b/2}\int_{\boldsymbol{b}_{\alpha 0}}\exp\left\{-\frac{1}{2}\boldsymbol{b}_{\alpha 0}^T\boldsymbol{b}_{\alpha 0}-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}\right.\\
&\qquad\left.\left.+C_1\Gamma\|\boldsymbol{b}_{\alpha 0}\|+C_2\Gamma\|\boldsymbol{Y}_i\|+C_3\Gamma\right\}d\boldsymbol{b}_{\alpha 0}\right]\Bigg]\\
=\ & C_7+\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left(\sum_{s=1}^{S} I(S_i=s)\widehat{\zeta}_s\right)+\frac{1}{2}\log(2\pi)\\
&+\frac{1}{n}\sum_{i=1}^{n} I(V_i=\tau)\left[\Gamma\log\Gamma-\Gamma\sum_{s=1}^{S} I(S_i=s)\widehat{\zeta}_s\right.\\
&\quad\left.+\log\mathrm{E}_{\alpha,\boldsymbol{b}_0}\left[\exp\left\{\frac{1}{2}\boldsymbol{M}_{i\alpha}^T\boldsymbol{M}_{i\alpha}-\frac{1}{2}\widehat{\boldsymbol{\mu}}_\alpha^T\widehat{\boldsymbol{\Sigma}}_b^{-1}\widehat{\boldsymbol{\mu}}_\alpha-\sum_{j=1}^{n_i}\frac{B(\widehat{\boldsymbol{\beta}};\boldsymbol{b}_{\alpha 0})}{A(D_i(t_j;\widehat{\phi}))}\right.\right.\right.
\end{aligned}$$

$$+C_1\Gamma\|\boldsymbol{b}_{\alpha 0}\| + C_2\Gamma\|\boldsymbol{Y}_i\| + C_3\Gamma\Big\}\Big]\Big]$$

$$= \quad C_8 + \frac{1}{n}\sum_{i=1}^{n}\Delta_i\Big(\sum_{s=1}^{S}\widehat{\zeta}_s\Big) - \frac{\Gamma}{n}\sum_{i=1}^{n}I(V_i=\tau)\Big(\sum_{s=1}^{S}\widehat{\zeta}_s\Big) + C_9(\Gamma), \tag{4.10}$$

where C_8 is a constant and $C_9(\Gamma)$ is a deterministic function of Γ. For the s-th stratum, (4.10) is that

$$0 \quad \leq \quad C_8 + \sum_{i=1}^{n}\Delta_i I(S_i=s)\widehat{\zeta}_s - \frac{\Gamma}{n}\sum_{i=1}^{n}I(V_i=\tau)I(S_i=s)\widehat{\zeta}_s + C_9(\Gamma).$$

By the strong law of large numbers, $\sum_{i=1}^{n}I(V_i=\tau)I(S_i=s)/n \longrightarrow P(V_i=\tau,S_i=s)>0$. Then, we can choose Γ large enough such that $\sum_{i=1}^{n}\Delta_i I(S_i=s)/n \leq (\Gamma/2n)\sum_{i=1}^{n}I(V_i=\tau)I(S_i=s)$. Thus, we obtain that

$$0 \quad \leq \quad C_8 + C_9(\Gamma) - \frac{\Gamma}{2n}\sum_{i=1}^{n}I(V_i=\tau)I(S_i=s)\widehat{\zeta}_s.$$

In other words,

$$\widehat{\zeta}_s \quad \leq \quad \frac{(C_8+C_9(\Gamma))2n}{\Gamma\sum_{i=1}^{n}I(V_i=\tau)I(S_i=s)} \quad \longrightarrow \quad \frac{(C_8+C_9(\Gamma))2}{\Gamma P(V_i=\tau,S_i=s)}.$$

If we denote $B_{s0}=\exp\big\{2\big(C_8+C_9(\Gamma)\big)/\big(\Gamma P(V_i=\tau,S_i=s)\big)\big\}$, we conclude that $\widehat{\Lambda}_s(\tau)\leq B_{s0}, s=1,\ldots,S$. Note that the above arguments hold for every sample in the probability space except a set with zero probability. Therefore, we have shown that, with probability one, $\widehat{\Lambda}_s(\tau)$ is bounded for any sample size n.

In the third step, the goal of this step is to show that, if $\widehat{\boldsymbol{\theta}}\to\boldsymbol{\theta}^*$ and $\widehat{\Lambda}_s$ weakly converges to Λ_s^* with probability one, then $\boldsymbol{\theta}^*=\boldsymbol{\theta}_0$ and $\Lambda_s^*=\Lambda_{s0}$, $s=1,\ldots,S$. We set some preliminaries as the followings: For convenience, we omit the index i for subject and use $\boldsymbol{O}$ to abbreviate the observed statistics $(\boldsymbol{Y},\boldsymbol{X},\widetilde{\boldsymbol{X}},V,\Delta,n_N,s)$ and $\{\boldsymbol{Z}(t),\widetilde{\boldsymbol{Z}}(t),0\leq t\leq V\}$ for a subject. By dropping $(\lambda_s(V))^{\Delta}$ from the complete data

likelihood function, we define

$$
\begin{aligned}
&G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\\
&=\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_i\boldsymbol{b}_\alpha)-B(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))}+C(Y_j;D(t_j;\phi))\right]\right\}\\
&\times\exp\left\{\Delta\big[\widetilde{\boldsymbol{Z}}(V)(\boldsymbol{\psi}\circ\boldsymbol{b}_\alpha)+\boldsymbol{Z}(V)\boldsymbol{\gamma}\big]-\int_0^V\exp\big\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}\circ\boldsymbol{b}_\alpha)+\boldsymbol{Z}(t)\boldsymbol{\gamma}\big\}d\Lambda_s(t)\right\}\\
&\times(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_\alpha)^T\boldsymbol{\Sigma}_b^{-1}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_\alpha)\right\}w_\alpha,
\end{aligned}
$$

and $$Q(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)=\frac{\sum_\alpha\int_{\boldsymbol{b}_\alpha}G(\boldsymbol{b}_\alpha,\alpha,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\exp\big\{\widetilde{\boldsymbol{Z}}(v)(\boldsymbol{\psi}\circ\boldsymbol{b}_\alpha)+\boldsymbol{Z}(v)\boldsymbol{\gamma}\big\}d\boldsymbol{b}_\alpha}{\sum_\alpha\int_{\boldsymbol{b}_\alpha}G(\boldsymbol{b}_\alpha,\alpha,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)d\boldsymbol{b}_\alpha}.$$

Furthermore, for any measurable function $f(\boldsymbol{O})$, we use operator notation to define $\mathbf{P}_n f=n^{-1}\sum_{i=1}^n f(\boldsymbol{O}_i)$ and $\mathbf{P}f=\int f d\mathbf{P}=\mathrm{E}[f(\boldsymbol{O})]$. Thus, $\mathbf{P}_n f$ is the empirical measure from n i.i.d observations and $\sqrt{n}(\mathbf{P}_n-\mathbf{P})$ is the empirical process based on these observations. We also define a class $\mathscr{F}=\{Q(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s):v\in[0,\tau],\boldsymbol{\theta}\in\Theta,\Lambda_s\in\mathbb{W},\Lambda_s(0)=0,\Lambda_s(\tau)\le B_{s0}\}$, where B_{s0} is the constant given in the second step and $\mathbb{W}$ contains all nondecreasing functions in $[0,\tau]$. According to the result proved in Section 4.4.3.1, $\mathscr{F}$ is P-Donsker.

Let m_s denote the number of subjects in stratum s; i.e. $n=\sum_{s=1}^S m_s$. V_s and Δ_s denote the observed time and censoring indicator for a subject belonging to stratum s, respectively. Thus, V_{sl} and Δ_{sl} are the l-th subject observed time and censoring indicator in stratum s.

Now we start the proof of the third step. Since $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\Lambda}})$ maximizes the function $l_n(\boldsymbol{\theta},\boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}=(\Lambda_1,\ldots,\Lambda_S)^T$ and Λ_s, $s=1,\ldots,S$, are any step functions with jumps only at V_i belonging to stratum s for which $\Delta_i=1$, we differentiate $l_n(\boldsymbol{\theta},\boldsymbol{\Lambda})$ with respect to $\Lambda_s\{V_{sl}\}$ and obtain the following equation, satisfied by $\widehat{\Lambda}_s$,

$$\widehat{\Lambda}_s\{V_{sl}\}=\frac{\Delta_{sl}}{m_s\,\mathbf{P}_{m_s}\big\{I(V_s\ge v)Q(v,\boldsymbol{O};\widehat{\boldsymbol{\theta}},\widehat{\Lambda}_s)\big\}\big|_{v=V_{sl}}}$$

Imitating the above equation, we also can construct another function, denoted by $\bar{\mathbf{\Lambda}} = (\bar{\Lambda}_1, \ldots, \bar{\Lambda}_S)^T$ such that $\bar{\Lambda}_s$, $s = 1, \ldots, S$, are also step functions with jumps only at the observed V_{sl} and the jump size $\bar{\Lambda}_s\{V_{sl}\}$ is given by

$$\bar{\Lambda}_s\{V_{sl}\} = \frac{\Delta_{sl}}{m_s \mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_{sl}}}.$$

Equivalently,

$$\bar{\Lambda}_s(t) = \frac{1}{m_s}\sum_{l=1}^{m_s} \frac{I(V_{sl} \leq t)\Delta_{sl}}{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_{sl}}}.$$

Then, we claim $\bar{\Lambda}_s(t)$ uniformly converges to $\Lambda_{s0}(t)$ in $[0, \tau]$. To prove the claim, note that

$$\begin{aligned}
&\sup_{t\in[0,\tau]}\left|\bar{\Lambda}_s(t) - \mathrm{E}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right]\right| \\
= &\sup_{t\in[0,\tau]}\left|\frac{1}{m_s}\sum_{l=1}^{m_s}\frac{I(V_{sl} \leq t)\Delta_{sl}}{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_{sl}}}\right. \\
&\qquad\qquad - \mathbf{P}_{m_s}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right] \\
&\left. + \mathbf{P}_{m_s}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right] - \mathbf{P}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right]\right| \\
\leq &\sup_{t\in[0,\tau]}\left|\frac{1}{m_s}\sum_{l=1}^{m_s} I(V_{sl} \leq t)\Delta_{sl}\left[\frac{1}{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}}\right.\right. \\
&\qquad\qquad \left.\left. - \frac{1}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}}\right]\right|_{v=V_{sl}}\Bigg| \\
&+ \sup_{t\in[0,\tau]}\left|\left(\mathbf{P}_{m_s} - \mathbf{P}\right)\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right]\right| \\
\leq &\sup_{t\in[0,\tau]}\left|\frac{1}{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}} - \frac{1}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}}\right| \\
&+ \sup_{t\in[0,\tau]}\left|\left(\mathbf{P}_{m_s} - \mathbf{P}\right)\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right]\right|. \qquad (4.11)
\end{aligned}$$

In (4.11), the right hand side converges to 0 because the first and second terms on the right hand side converges to 0 in the following: First, according to Section 4.4.3.1, $\{Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) : v \in [0, \tau]\}$ is a bounded and Glivenko-Cantelli class. $\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) : v \in [0, \tau]\}$ is also a Glivenko-Cantelli class because $\{I(V_s \geq v) : v \in [0, \tau]\}$ is a Glivenko-Cantelli class and the functional $(f, g) \to fg$ for any bounded two functions f and g is Lipschitz continuous. Then, we obtain that $\sup_{t \in [0,\tau]} \left| \mathbf{P}_{m_s}\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\} - \mathbf{P}\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\} \right|$ converges to 0. Besides, from Section 4.4.3.1, $\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\} > \mathbf{P}\left\{I(V_s \geq v) \exp\{-C_{10} - C_{11}\|\boldsymbol{Y}\|\}\right\}$ for the two constants C_{10} and C_{11}, which means $\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}$ is bounded from below. Thus, the first term tends to 0. Second, since the class $\left\{I(V_s \leq t)\Delta_s / \mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s} : t \in [0, \tau]\right\}$ is also a Glivenko-Cantelli class, the second term vanishes as m_s goes to infinity.

Therefore, we conclude that $\bar{\Lambda}_s(t)$ uniformly converges to

$$\mathrm{E}\left[\frac{I(V_s \leq t)\Delta_s}{\mathbf{P}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}\big|_{v=V_s}}\right]. \tag{4.12}$$

We can easily verify that (4.12) is equal to $\Lambda_{s0}(t)$. Thus, the claim that $\bar{\Lambda}_s(t)$ uniformly converges to $\Lambda_{s0}(t)$ in $[0, \tau]$ has been proved.

From the construction of $\bar{\Lambda}_s(t)$, we obtain

$$\widehat{\Lambda}_s(t) = \int_0^t \frac{d\widehat{\Lambda}_s(v)}{d\bar{\Lambda}_s(v)} d\bar{\Lambda}_s(v) = \int_0^t \frac{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}}{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)\right\}} d\bar{\Lambda}_s(v). \tag{4.13}$$

$\widehat{\Lambda}_s(t)$ is absolutely continuous with respect to $\bar{\Lambda}_s(t)$. On the other hand, since both $\{I(V_s \geq v) : v \in [0, \tau]\}$ and $\mathscr{F}$ are Glivenko-Cantelli classes, $\{I(V_s \geq v)Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) : v \in [0, \tau]\}$ is also a Glivenko-Cantelli class. Thus, we have

$$\sup_{v\in[0,\tau]}\left|(\mathbf{P}_{m_s}-\mathbf{P})\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\widehat{\boldsymbol{\theta}},\widehat{\Lambda}_s)\right\}\right|+\sup_{v\in[0,\tau]}\left|(\mathbf{P}_{m_s}-\mathbf{P})\left\{I(V_s\geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}\right|$$
$$\longrightarrow 0 \quad \text{a.s.}$$

By the bounded convergence theorem and the fact that $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ and $\widehat{\Lambda}_s$ converges to Λ_s^*, for each v, $\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\widehat{\boldsymbol{\theta}},\widehat{\Lambda}_s)\right\} \longrightarrow \mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)\right\}$; moreover, it is straightforward to check the derivative of $\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\widehat{\boldsymbol{\theta}},\widehat{\Lambda}_s)\right\}$ with respect to v. Thus, by the Arzela-Ascoli theorem, uniformly in $[0,\tau]$,

$$\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\widehat{\boldsymbol{\theta}},\widehat{\Lambda}_s)\right\} \longrightarrow \mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)\right\}.$$

Then, combining the above result and (4.13), it holds that, uniformly in $[0,\tau]$,

$$\frac{d\widehat{\Lambda}_s(v)}{d\bar{\Lambda}_s(v)} = \frac{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}}{\mathbf{P}_{m_s}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\widehat{\boldsymbol{\theta}},\widehat{\Lambda}_s)\right\}} \longrightarrow \frac{\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}}{\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)\right\}}. \quad (4.14)$$

After taking limits on both sides of (4.13), we obtain that

$$\lim_{m_s\to\infty}\widehat{\Lambda}_s(t) = \int_0^t \frac{\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\right\}}{\mathbf{P}\left\{I(V_s \geq v)Q(v,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)\right\}} d\Lambda_{s0}(v), \quad (4.15)$$

Therefore, since $\Lambda_{s0}(t)$ is differentiable with respect to the Lebesque measure, so is $\Lambda_s^*(t)$; that is, (4.15) is equal to

$$\int_0^t \frac{d\Lambda_s^*(v)}{d\Lambda_{s0}(v)} d\Lambda_{s0}(v). \quad (4.16)$$

And we denote $\lambda_s^*(t)$ as the derivative of $\Lambda_s^*(t)$. Additionally, from (4.14) ~ (4.16), note that $\widehat{\Lambda}_s\{V_s\}/\bar{\Lambda}_s\{V_s\}$ uniformaly converges to $d\Lambda_s^*(V_s)/d\Lambda_{s0}(V_s) = \lambda_s^*(V_s)/\lambda_{s0}(V_s)$. Therefore, a second conclusion is that $\widehat{\Lambda}_s$ uniformly converges to Λ_s^* since Λ_s^* is contin-

uous.

On the other hand,

$$
\begin{aligned}
& n^{-1} l_n(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}) - n^{-1} l_n(\boldsymbol{\theta}_0, \bar{\boldsymbol{\Lambda}}) \\
&= \sum_{s=1}^{S} \left(\mathbf{P}_{m_s} \left[\Delta_s \log \frac{\widehat{\Lambda}_s\{V_s\}}{\bar{\Lambda}_s\{V_s\}} \right] + \mathbf{P}_{m_s} \left[\log \frac{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s) d\boldsymbol{b}_\alpha}{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \bar{\Lambda}_s) d\boldsymbol{b}_\alpha} \right] \right) \\
&\geq 0. \qquad (4.17)
\end{aligned}
$$

Using the result of Section 4.4.3.1 and similar arguments as above, we can verify that

$$
\log \frac{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s) d\boldsymbol{b}_\alpha}{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \bar{\Lambda}_s) d\boldsymbol{b}_\alpha}
$$

belongs to a Glivenko-Cantelli class and

$$
\mathbf{P} \left[\log \frac{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s) d\boldsymbol{b}_\alpha}{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \bar{\Lambda}_s) d\boldsymbol{b}_\alpha} \right] \longrightarrow \mathbf{P} \left[\log \frac{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha}{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha} \right].
$$

Since $\widehat{\Lambda}_s\{V_s\}/\bar{\Lambda}_s\{V_s\}$ uniformaly converges to $\lambda_s^*(V_s)/\lambda_{s0}(V_s)$, we obtain that, from (4.17),

$$
\mathbf{P} \left[\log \left\{ \frac{\left(\lambda_s^*(V_s)\right)^{\Delta_s} \sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha}{\left(\lambda_{s0}(V_s)\right)^{\Delta_s} \sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha} \right\} \right] \geq 0.
$$

Note that the left-hand side of the inequality is the negative Kullback-Leibler information. Then, the equality holds with probability one, and it immediately follows

$$
\left(\lambda_s^*(V_s)\right)^{\Delta_s} \sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha = \left(\lambda_{s0}(V_s)\right)^{\Delta_s} \sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha. \qquad (4.18)
$$

Our proof will be completed if we can show $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ and $\Lambda_s^* = \Lambda_{s0}$ from (4.18). Since (4.18) holds with probability one, (4.18) holds for any $(V_s, \Delta_s = 1)$ and the case $(V_s = \tau, \Delta_s = 0)$, but may not hold for $(V_s, \Delta_s = 0)$ when $V_s \in (0, \tau)$. However, we can show

that (4.18) is also true for $(V_s, \Delta_s = 0)$ when $V_s \in (0, \tau)$. To do this, treating both sides of (4.18) as functions of V_s, we integrate these functions over an interval (V_s, τ) for $\Delta_s = 0$ as the following;

$$\int_{V_s}^{\tau} \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha \quad = \quad \int_{V_s}^{\tau} \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha$$

to obtain that

$$\begin{aligned}
&\sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0, V_s=\tau} - \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0, V_s=V_s} \\
&= \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0, V_s=\tau} - \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0, V_s=V_s}.
\end{aligned}$$

After comparing this above equality with another following equality, which is given by (4.18) at $\Delta_s = 0$ and $V_s = \tau$,

$$\sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0, V_s=\tau} \quad = \quad \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0, V_s=\tau},$$

we obtain

$$\sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0, V_s=V_s} \quad = \quad \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0, V_s=V_s},$$

and therefore

$$\sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0} \quad = \quad \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\boldsymbol{b}_\alpha \Big|_{\Delta_s=0};$$

that is, (4.18) also holds for any V_s and $\Delta_s = 0$.

Thus, first to show that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$, $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{b0}$, $\boldsymbol{\mu}_\alpha^* = \boldsymbol{\mu}_{\alpha 0}$, and $w_\alpha^* = w_{\alpha 0}$, $\alpha = 1, \ldots, K$, we let $\Delta_s = 0$ and $V_s = 0$ in (4.18). After integrating over $\boldsymbol{b}_\alpha$ and summing

up over α, we have that, with probability one,

$$
\begin{aligned}
&\sum_{\alpha}\int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)d\boldsymbol{b}_\alpha\Big|_{\Delta_s=0,V_s=0} = \sum_{\alpha}\int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})d\boldsymbol{b}_\alpha\Big|_{\Delta_s=0,V_s=0} \\
\Rightarrow\ &\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}^*+\widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha)-B(\boldsymbol{\beta}^*,\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}+C(Y_j;D(t_j;\phi^*))\right]\right\} \\
&\qquad\times(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b^*|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_\alpha^*)^T\boldsymbol{\Sigma}_b^{*-1}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_\alpha^*)\right\}d\boldsymbol{b}_\alpha \\
&=\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}_0+\widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha)-B(\boldsymbol{\beta}_0,\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}+C(Y_j;D(t_j;\phi_0))\right]\right\} \\
&\qquad\times(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_{b0}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_{\alpha0})^T\boldsymbol{\Sigma}_{b0}^{-1}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_{\alpha0})\right\}d\boldsymbol{b}_\alpha \\
\Rightarrow\ &\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))}+C(Y_j;D(t_j;\phi^*))\right]\right\}(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b^*|^{-1/2} \\
&\times\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha}{A(D(t_j;\phi^*))}-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}\right. \\
&\qquad\left.-\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_b^{*-1}\boldsymbol{b}_\alpha+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\boldsymbol{b}_\alpha-\frac{1}{2}\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\boldsymbol{\mu}_\alpha^*\right\}w_\alpha^*d\boldsymbol{b}_\alpha \\
&=\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}_0}{A(D(t_j;\phi_0))}+C(Y_j;D(t_j;\phi_0))\right]\right\}(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_{b0}|^{-1/2} \\
&\times\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha}{A(D(t_j;\phi_0))}-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\right. \\
&\qquad\left.-\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}_\alpha+\boldsymbol{\mu}_{\alpha0}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}_\alpha-\frac{1}{2}\boldsymbol{\mu}_{\alpha0}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{\mu}_{\alpha0}\right\}w_{\alpha0}d\boldsymbol{b}_\alpha.
\end{aligned}
$$

By some algebra, the left hand side becomes

$$
\begin{aligned}
&\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))}+C(Y_j;D(t_j;\phi^*))\right]\right\}(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b^*|^{-1/2} \\
&\times\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{-\frac{1}{2}\left[\left(\Sigma_b^{*-1/2}\boldsymbol{b}_\alpha\right)^T\left(\Sigma_b^{*-1/2}\boldsymbol{b}_\alpha\right)\right.\right. \\
&-2\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)\boldsymbol{\Sigma}_b^{*1/2}\boldsymbol{\Sigma}_b^{*-1/2}\boldsymbol{b}_\alpha \\
&\left.\left.+\left[\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)\boldsymbol{\Sigma}_b^{*1/2}\right]\left[\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)\boldsymbol{\Sigma}_b^{*1/2}\right]^T\right]\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{2}\Bigg[\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)\boldsymbol{\Sigma}_b^{*1/2}\Bigg]\Bigg[\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)\boldsymbol{\Sigma}_b^{*1/2}\Bigg]^T\\
&-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}-\frac{1}{2}\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\boldsymbol{\mu}_\alpha^*\Bigg\}w_\alpha^*d\boldsymbol{b}_\alpha\\
=&\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))}+C(Y_j;D(t_j;\phi^*))\right]\right\}(2\pi)^{-d_b/2}|\boldsymbol{\Sigma}_b^*|^{-1/2}\\
&\times\sum_\alpha w_\alpha^*\Bigg[\exp\left\{\frac{1}{2}\Bigg[\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)\boldsymbol{\Sigma}_b^{*1/2}\Bigg]\right.\\
&\qquad\left.\times\Bigg[\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)\boldsymbol{\Sigma}_b^{*1/2}\Bigg]^T-\frac{1}{2}\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\boldsymbol{\mu}_\alpha^*\right\}\\
&\quad\times\int_{\boldsymbol{b}_\alpha}\exp\left\{-\frac{1}{2}\Bigg[\Sigma_b^{*-1/2}\boldsymbol{b}_\alpha-\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)^T\Bigg]^T\right.\\
&\qquad\left.\times\Bigg[\Sigma_b^{*-1/2}\boldsymbol{b}_\alpha-\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)^T\Bigg]\right\}\\
&\quad\times\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}\right\}d\boldsymbol{b}_\alpha\Bigg]\\
=&\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))}+C(Y_j;D(t_j;\phi^*))\right]\right\}\\
&\times\sum_\alpha w_\alpha^*\Bigg[\exp\Bigg\{\frac{1}{2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)\boldsymbol{\Sigma}_b^*\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}+\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)^T\\
&\qquad-\frac{1}{2}\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\boldsymbol{\mu}_\alpha^*\Bigg\}\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N}\left(\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}\right)\right\}\right]\Bigg]\\
=&\exp\left\{\sum_{j=1}^{n_N}\left[\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}^*}{A(D(t_j;\phi^*))}+C(Y_j;D(t_j;\phi^*))\right]\right\}\\
&\times\sum_\alpha w_\alpha^*\Bigg[\exp\left\{\frac{1}{2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)\boldsymbol{\Sigma}_b^*\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)^T\right.\\
&\quad\left.+\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)\boldsymbol{\Sigma}_b^*\left(\boldsymbol{\mu}_\alpha^{*T}\boldsymbol{\Sigma}_b^{*-1}\right)^T\right\}\times\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}\right\}\right]\Bigg]\\
=&\exp\left\{\sum_{j=1}^{n_N}C(Y_j;D(t_j;\phi^*))\right\}\sum_\alpha w_\alpha^*\Bigg[\exp\Bigg\{\frac{1}{2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)\boldsymbol{\Sigma}_b^*\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)^T\\
&\qquad+\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi^*))}\left(\boldsymbol{X}_j\boldsymbol{\beta}^*+\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_\alpha^*\right)\Bigg\}\times\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}\right\}\right]\Bigg]\\
=&\sum_\alpha\Bigg[\exp\left\{\frac{1}{2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)\boldsymbol{\Sigma}_b^*\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)^T\right.
\end{aligned}
$$

$$+\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi^*))}\left(\boldsymbol{X}_j\boldsymbol{\beta}^*+\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_\alpha^*\right)\Bigg\}$$
$$\times\exp\left\{\sum_{j=1}^{n_N}C(Y_j;D(t_j;\phi^*))\right\}w_\alpha^*\,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}\right\}\right]\Bigg]. \quad (4.19)$$

Likewise, the right-hand side becomes

$$\sum_\alpha\Bigg[\exp\left\{\frac{1}{2}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\right)\boldsymbol{\Sigma}_{b0}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\right)^T\right.$$
$$\left.+\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\left(\boldsymbol{X}_j\boldsymbol{\beta}_0+\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha0}\right)\right\}$$
$$\times\exp\left\{\sum_{j=1}^{n_N}C(Y_j;D(t_j;\phi_0))\right\}w_{\alpha0}\,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\right\}\right]\Bigg]. \quad (4.20)$$

Then, to compare the coefficients of $\boldsymbol{Y}^T\boldsymbol{Y}$ and $\boldsymbol{Y}$ in the exponential part and the constant term out of the exponential part from (4.19) and (4.20), we have

$$\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)\boldsymbol{\Sigma}_b^*\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi^*))}\right)^T=\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\right)\boldsymbol{\Sigma}_{b0}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\right)^T, \quad (4.21)$$

$$\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi^*))}\left(\boldsymbol{X}_j\boldsymbol{\beta}^*+\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_\alpha^*\right)=\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\left(\boldsymbol{X}_j\boldsymbol{\beta}_0+\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha0}\right), \quad (4.22)$$

and

$$\exp\left\{\sum_{j=1}^{n_N}C(Y_j;D(t_j;\phi^*))\right\}w_\alpha^*\,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}\right\}\right]$$
$$=\exp\left\{\sum_{j=1}^{n_N}C(Y_j;D(t_j;\phi_0))\right\}w_{\alpha0}\,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\right\}\right]. \quad (4.23)$$

Furthermore, by the assumption of the generalized linear mixed model with canonical link function for longitudinal outcome $Y(t)$ at time t, we have $\mu(t)=\mathrm{E}(Y(t)|\boldsymbol{b})=B'(\eta(t))$ and $v(t)=\mathrm{Var}(Y(t)|\boldsymbol{b})=B''(\eta(t))A(\phi(t))$, where $\boldsymbol{b}=\sum_{k=1}^K I(\alpha=k)\boldsymbol{b}_k$, $\eta(t)=g(\mu(t))=\boldsymbol{X}(t)\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}(t)\boldsymbol{b}$, $v(t)=v(\mu(t))A(\phi(t))$, $g(\cdot)$ and $v(\cdot)$ are known link and

variance functions respectively, and $B'(\eta(t))$ and $B''(\eta(t))$ are the first and second derivatives of $B(\eta(t))$ with respect to the canonical parameter $\eta(t)$. Hence, we have

$$\mathrm{E}(Y_j|\boldsymbol{b}) = B'(\eta_j) = B'(\boldsymbol{\beta}^*;\boldsymbol{b}) = B'(\boldsymbol{\beta}_0;\boldsymbol{b}) \tag{4.24}$$

and

$$\mathrm{Var}(Y_j|\boldsymbol{b}) = B''(\eta_j)A(D(t_j;\phi)) = B''(\boldsymbol{\beta}^*;\boldsymbol{b})A(D(t_j;\phi^*)) = B''(\boldsymbol{\beta}_0;\boldsymbol{b})A(D(t_j;\phi_0)). \tag{4.25}$$

By the continuous mapping theorem and (4.24), we obtain $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. Then, (4.25) becomes $B''(\boldsymbol{\beta}_0;\boldsymbol{b})A(D(t_j;\phi^*)) = B''(\boldsymbol{\beta}_0;\boldsymbol{b})A(D(t_j;\phi_0))$. Hence, by assumption (A6), $A(D(t_j;\phi^*)) = A(D(t_j;\phi_0))$, and, by the continuous mapping theorem, we obtain $D(t_j;\phi^*) = D(t_j;\phi_0)$, $j = 1,\ldots,n_N$, and $\phi^* = \phi_0$. Thus, (4.21) can be written as

$$\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\right)\boldsymbol{\Sigma}_b^*\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\right)^T = \left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\right)\boldsymbol{\Sigma}_{b0}\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}\right)^T.$$

Then, by assumption (A6), we obtain $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{b0}$. Since $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\phi^* = \phi_0$, (4.22) can be written as

$$\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\left(\boldsymbol{X}_j\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_\alpha^*\right) = \sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\left(\boldsymbol{X}_j\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\right).$$

Also, by assumption (A6), we obtain $\boldsymbol{\mu}_\alpha^* = \boldsymbol{\mu}_{\alpha 0}$, $\alpha = 1,\ldots,K$. In (4.23) for the constant terms, note that the random effect $\boldsymbol{b}_\alpha$ on the left-hand side follows a multivariate normal distribution with mean $\boldsymbol{\Sigma}_b^*\big(\sum_{j=1}^{n_N} Y_j\widetilde{\boldsymbol{X}}_j/A(D(t_j;\phi^*))\big)^T + \boldsymbol{\mu}_\alpha^*$ and covariance $\boldsymbol{\Sigma}_b^*$ and the random effect $\boldsymbol{b}_\alpha$ on the right-hand side follows a multivariate normal distribution with mean $\boldsymbol{\Sigma}_{b0}\big(\sum_{j=1}^{n_N} Y_j\widetilde{\boldsymbol{X}}_j/A(D(t_j;\phi_0))\big)^T + \boldsymbol{\mu}_{\alpha 0}$ and covariance $\boldsymbol{\Sigma}_{b0}$. (i) Because $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{b0}$, $\boldsymbol{\mu}_\alpha^* = \boldsymbol{\mu}_{\alpha 0}$, $\alpha = 1,\ldots,K$, and $\phi^* = \phi_0$, the random effects $\boldsymbol{b}_\alpha$'s on both sides

follow the same multivariate normal distribution. (ii) Besides, because $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\phi^* = \phi_0$, we have $\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))} = \sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}$. By (i) and (ii), we obtain (iii) $\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\big[\exp\big\{\frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))}\big\}\big] = \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\big[\exp\big\{\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\big\}\big]$. Also, (iv) since $\phi^* = \phi_0$, we have $\exp\big\{\sum_{j=1}^{n_N} C(Y_j; D(t_j;\phi^*))\big\} = \exp\big\{\sum_{j=1}^{n_N} C(Y_j; D(t_j;\phi_0))\big\}$. By (iii) and (iv), (4.23) can be written as

$$\exp\left\{\sum_{j=1}^{n_N} C(Y_j; D(t_j;\phi_0))\right\} w_\alpha^* \,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\right\}\right]$$
$$= \exp\left\{\sum_{j=1}^{n_N} C(Y_j; D(t_j;\phi_0))\right\} w_{\alpha 0} \,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\right\}\right].$$

Then, by assumption (A6), we obtain $w_\alpha^* = w_{\alpha 0}$. $\alpha = 1, \ldots, K$.

Next, to show that $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ and $\Lambda_s^* = \Lambda_{s0}$, we let $\Delta_s = 0$ in (4.18). Through the similar arguments done for the proof of $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$, $\boldsymbol{\Sigma}_b^* = \boldsymbol{\Sigma}_{b0}$, $\boldsymbol{\mu}_\alpha^* = \boldsymbol{\mu}_{\alpha 0}$, and $w_\alpha^* = w_{\alpha 0}$, $\alpha = 1, \ldots, K$, we obtain

$$\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))} - \int_0^{V_s} \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}^* \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}^*\right\} d\Lambda_s^*(t)\right\}\right]$$
$$= \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} - \int_0^{V_s} \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0 \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\} d\Lambda_{s0}(t)\right\}\right], \tag{4.26}$$

where the random effects $\boldsymbol{b}_\alpha$'s on both sides follow a multivariate normal distribution with mean $\boldsymbol{\Sigma}_{b0}\big(\sum_{j=1}^{n_N} Y_j \widetilde{\boldsymbol{X}}_j / A(D(t_j;\phi_0))\big)^T + \boldsymbol{\mu}_{\alpha 0}$ and covariance $\boldsymbol{\Sigma}_{b0}$.

For any fixed $\widetilde{\boldsymbol{X}} = \big(\widetilde{\boldsymbol{X}}_1^T, \ldots, \widetilde{\boldsymbol{X}}_{n_N}^T\big)^T$, treating $\widetilde{\boldsymbol{X}}^T\boldsymbol{Y}$ as a parameter in this normal family, $\boldsymbol{b}_\alpha = \sum_{k=1}^K I(\alpha = k)\boldsymbol{b}_k$ is the complete statistic for $\widetilde{\boldsymbol{X}}^T\boldsymbol{Y}$. Therefore,

$$\exp\left\{-\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}^*;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi^*))} - \int_0^{V_s} \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}^* \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}^*\right\} d\Lambda_s^*(t)\right\}$$
$$= \exp\left\{-\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} - \int_0^{V_s} \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0 \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\} d\Lambda_{s0}(t)\right\}.$$

Since $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$, equivalently, we have

$$\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}^* \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}^*\right\}\lambda_s^*(t) = \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0 \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}\lambda_{s0}(t).$$

By assumptions (A3) and (A6), $\boldsymbol{\psi}^* = \boldsymbol{\psi}_0$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ and $\Lambda_s^* = \Lambda_{s0}$.

Since all the three steps are completed, we can conclude that, with probability one, $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\Lambda}}$ converges to $\boldsymbol{\Lambda}_0$ in $[0, \tau]$. Moreover, as mentioned in the beginning of this proof for consistency, since $\boldsymbol{\Lambda}_0$ is continuous in $[0, \tau]$, the latter can be strengthened to uniform convergence; that is, $\sup_{t\in[0,\tau]} ||\widehat{\boldsymbol{\Lambda}}(t) - \boldsymbol{\Lambda}_0(t)|| \to 0$ almost surely. Therefore, Theorem 4.1 is proved.

4.4.2 Proof of asymptotic normality

Asymptotic distribution for the proposed estimator can be shown if we can verify the conditions of Theorem 3.3.1 (p310) in van der Vaart and Wellner (1996). Then, we will show that the distribution is normal. For completeness, we state this theorem below following Theorem 4 in Appendix A of Parner (1998).

Theorem 4.3. (Theorem 3.3.1 in van der Vaart and Wellner, 1996; Theorem 4 in Parner, 1998) *Let U_n and U be random maps and a fixed map, respectively, from ξ to a Banach space such that:*

(a) $\sqrt{n}(U_n - U)(\widehat{\xi}_n) - \sqrt{n}(U_n - U)(\xi_0) = o_P^*(1 + \sqrt{n}||\widehat{\xi}_n - \xi_0||)$.

(b) The sequence $\sqrt{n}(U_n - U)(\xi_0)$ converges in distribution to a tight random element $\boldsymbol{Z}$.

(c) the function $\xi \to U(\xi)$ is Fréchet differentiable at ξ_0 with a continuously invertible derivative ∇U_{ξ_0} (on its range).

(d) U_{ξ_0} and $\widehat{\xi}_n$ satisfies $U_n(\widehat{\xi}_n) = o_P^*(n^{-1/2})$ and converges in outer probability to ξ_0.

Then $\sqrt{n}(\widehat{\xi}_n - \xi_0) \Rightarrow \nabla U_{\xi_0}^{-1} \boldsymbol{Z}$.

We will prove the conditions (a)~(d). In our situation, the parameter $\xi_s = (\boldsymbol{\theta}, \boldsymbol{\Lambda}_s) \in \Xi = \{(\boldsymbol{\theta}, \boldsymbol{\Lambda}_s) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda_s(t) - \Lambda_{s0}(t)| \le \delta,\ s = 1, \ldots, S\}$ for a fixed small constant δ. We note that Ξ is a convex set. Define a set $\mathscr{H} = \{(\mathbf{h}_1, h_2) : \|\mathbf{h}_1\| \le 1, \|h_2\|_V \le 1\}$, where $\|h_2\|_V$ is the total variation of h_2 in $[0, \tau]$ defined as

$$\sup_{0=t_0\le t_1\le\cdots\le t_l=\tau} \sum_{j=1}^{l} |h_2(t_j) - h_2(t_{j-1})|.$$

Furthermore, we define that, for stratum s,

$$\begin{aligned} U_{m_s}(\xi_s)(\mathbf{h}_1, h_2) &= \mathbf{P}_{m_s}\{l_\theta(\boldsymbol{\theta}, \Lambda_s)^T \mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2]\} \\ \text{and} \qquad U_s(\xi_s)(\mathbf{h}_1, h_2) &= \mathbf{P}\{l_\theta(\boldsymbol{\theta}, \Lambda_s)^T \mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2]\}, \end{aligned}$$

where $l_\theta(\boldsymbol{\theta}, \Lambda_s)$ is the first derivative of the log-likelihood function from one single subject belonging to stratum s, denoted by $l(\boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)$, with respect to $\boldsymbol{\theta}$, and $l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)$ is the derivative of $l(\boldsymbol{O}; \boldsymbol{\theta}, \Lambda_{s\varepsilon})$ at $\varepsilon = 0$, where $\Lambda_{s\varepsilon}(t) = \int_0^t (1 + \varepsilon h_2(u)) d\Lambda_s(u)$. Therefore, we can see that both U_{m_s} and U_s map from Ξ to $\ell^\infty(\mathscr{H})$ and $\sqrt{m_s}\{U_{m_s}(\xi_s) - U_s(\xi_s)\}$ is an empirical process in the space $\ell^\infty(\mathscr{H})$.

Denote $(\mathbf{h}_1^\beta, \mathbf{h}_1^\phi, \mathbf{h}_1^{\Sigma_b}, \mathbf{h}_1^\mu, \mathbf{h}_1^w, \mathbf{h}_1^\psi, \mathbf{h}_1^\gamma)$ as the corresponding components of $\mathbf{h}_1$ for the parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathrm{Vec}(\boldsymbol{\Sigma}_b),\ \boldsymbol{\mu}, \boldsymbol{w}, \boldsymbol{\psi}, \boldsymbol{\gamma})$, respectively. From Section 4.4.3.2, for any $(\mathbf{h}_1, h_2) \in \mathscr{H}$, the class

$$\begin{aligned} \mathscr{G} = \Big\{ & l_\theta(\boldsymbol{\theta}, \Lambda_s)^T \mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T \mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2], \\ & \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda_s(t) - \Lambda_{s0}(t)| \le \delta,\ (\mathbf{h}_1, h_2) \in \mathscr{H} \Big\} \end{aligned}$$

is shown as P-Donsker (Section 2.1 of van der Vaart and Wellner (1996), and it is also

implied that

$$\sup_{(\mathbf{h}_1,h_2)\in\mathscr{H}} \mathbf{P}\left[l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0,\Lambda_{s0})^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}_0,\Lambda_{s0})[h_2]\right]^2 \longrightarrow 0$$

as $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]}|\Lambda_s(t)-\Lambda_{s0}(t)| \to 0$. Then we conclude the followings:

(a) follows from Lemma 3.3.5 (p311) of van der Vaart and Wellner (1996).

(b) holds as a result of Section 4.4.3.2 and the convergence is defined in the metric space $\ell^\infty(\mathscr{H})$ by the Donsker theorem (Section 2.5 of van der Vaart and Wellner (1996).

(d) is true because $(\widehat{\boldsymbol{\theta}},\widehat{\Lambda}_s)$ maximizes $\mathbf{P}_{m_s} l(\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)$, $(\boldsymbol{\theta}_0,\Lambda_{s0})$ maximizes $\mathbf{P}\, l(\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)$, and $(\widehat{\boldsymbol{\theta}},\widehat{\Lambda}_s)$ converges to $(\boldsymbol{\theta}_0,\Lambda_{s0})$ from Theorem 4.1.

Now, we need to verify the conditions in (c). Since the proof of the first half in (c), that the function $\xi \to U(\xi)$ is *Fréchet* differentiable at ξ_0, is given in Section 4.4.3.3, we will only prove that the derivative ∇U_{ξ_0} is continuously invertible on its range $\ell^\infty(\mathscr{H})$. According to Section 4.4.3.3, ∇U_{ξ_0} can be expressed as follows: for any $(\boldsymbol{\theta}_1,\Lambda_{s1})$ and $(\boldsymbol{\theta}_2,\Lambda_{s2})$ in Ξ,

$$\nabla U_{\xi_0}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\Lambda_{s1}-\Lambda_{s2})[\mathbf{h}_1,h_2] = (\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^T\Omega_1[\mathbf{h}_1,h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1,h_2]d(\Lambda_{s1}-\Lambda_{s2})(t), \tag{4.27}$$

where both Ω_1 and Ω_2 are linear operators on $\mathscr{H}$, and $\Omega = (\Omega_1,\Omega_2)$ maps $\mathscr{H} \subset \boldsymbol{R}^d \times \mathrm{BV}[0,\tau]$ to $\boldsymbol{R}^d \times \mathrm{BV}[0,\tau]$, where $\mathrm{BV}[0,\tau]$ contains all the functions with finite total variation in $[0,\tau]$. The explicit expressions of Ω_1 and Ω_2 are given in Section 4.4.3.3. From (4.27), we can treat $(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\Lambda_{s1}-\Lambda_{s2})$ as an element in $\ell^\infty(\mathscr{H})$ via the following definition:

$$(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\Lambda_{s1}-\Lambda_{s2})[\mathbf{h}_1,h_2] = (\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^T\mathbf{h}_1 + \int_0^\tau h_2(t)d(\Lambda_{s1}-\Lambda_{s2})(t), \quad \forall (\mathbf{h}_1,h_2)\in\boldsymbol{R}^d\times\mathrm{BV}[0,\tau].$$

Then ∇U_{ξ_0} can be expanded as a linear operator from $\ell^\infty(\mathscr{H})$ to itself. Therefore, if we can show that there exists some positive constant ε such that $\varepsilon\mathscr{H} \subset \Omega(\mathscr{H})$, then we will have that for any $(\delta\boldsymbol{\theta}, \delta\Lambda_s) \in \ell^\infty(\mathscr{H})$,

$$\begin{aligned}\|\nabla U_{\xi_0}(\delta\boldsymbol{\theta}, \delta\Lambda_s)\|_{\ell^\infty(\mathscr{H})} &= \sup_{(\mathbf{h}_1,h_2)\in\mathscr{H}} \left|\delta\boldsymbol{\theta}^T\Omega_1[\mathbf{h}_1,h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1,h_2]d\delta\Lambda_s(t)\right| \\ &= \|(\delta\boldsymbol{\theta}, \delta\Lambda_s)\|_{\ell^\infty(\Omega(\mathscr{H}))} \geq \varepsilon\|(\delta\boldsymbol{\theta}, \delta\Lambda_s)\|_{\ell^\infty(\mathscr{H})},\end{aligned}$$

and ∇U_{ξ_0} will be continuously invertible.

Note that to prove $\varepsilon\mathscr{H} \subset \Omega(\mathscr{H})$ for some ε is equivalent to showing that Ω is invertible. We also note from Section 4.4.3.3, that Ω is the summation of an invertible operator and a compact operator. By Theorem 4.25 of Rudin (1973), for the proof of the invertibility of Ω, it is sufficient to verify that Ω is one to one: if $\Omega[\mathbf{h}_1, h_2] = 0$, then, by choosing $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 = \varepsilon^*\mathbf{h}_1$ and $\Lambda_{s1} - \Lambda_{s2} = \varepsilon^*\int h_2 d\Lambda_{s0}$ in (4.27) for a small constant ε^*, we obtain

$$\nabla U_{\xi_0}(\mathbf{h}_1, \int h_2 d\Lambda_{s0})[\mathbf{h}_1, h_2] = \varepsilon^*(\mathbf{h}_1^T, h_2)\begin{pmatrix}\Omega_1[\mathbf{h}_1,h_2]\\ \Omega_2[\mathbf{h}_1,h_2]\end{pmatrix} = \varepsilon^*(\mathbf{h}_1^T, h_2)\Omega[\mathbf{h}_1,h_2] = 0.$$

By the definition of ∇U_{ξ_0}, we note that $\nabla U_{\xi_0}(\mathbf{h}_1, \int h_2 d\Lambda_{s0})[\mathbf{h}_1, h_2]$ is the negative information matrix in the submodel $(\boldsymbol{\theta}_0 + \varepsilon\mathbf{h}_1, \Lambda_{s0} + \varepsilon\int h_2 d\Lambda_{s0})$. Thus, the score function along this submodel should be zero with probability one; that is, $l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2] = 0$; that is, with probability one, for the numerator of the score function

$$\begin{aligned}0 = \sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \times \Bigg[&\sum_{j=1}^{n_N} \frac{1}{A(D(t_j;\phi_0))}\big(Y_j\boldsymbol{X}_j - B'(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)\big)\mathbf{h}_1^\beta \\ &+ \sum_{j=1}^{n_N}\left\{-\left(\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha) - B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))^2}\right)A'(D(t_j;\phi_0)) + C'(Y_j; D(t_j;\phi_0))\right\} \mathrm{h}_1^\phi\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_{\alpha 0})^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_{\alpha 0})-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\right)\\
&+\left(\boldsymbol{b}_\alpha-\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}\right)^T\boldsymbol{\Sigma}_{b0}^{-1}\mathbf{h}_1^{\mu_\alpha}+\frac{1}{w_{\alpha 0}}\mathbf{h}_1^{w_{\alpha 0}}+\Delta_s\left\{\left(\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{b}_\alpha^T\right)\mathbf{h}_1^{\psi}+\boldsymbol{Z}(V_s)\mathbf{h}_1^{\gamma}\right\}\\
&-\int_0^{V_s}\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b}_\alpha)+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}\times\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{b}_\alpha^T)\mathbf{h}_1^{\psi}+\boldsymbol{Z}(t)\mathbf{h}_1^{\gamma}\right\}d\Lambda_{s0}(t)\Bigg]d\boldsymbol{b}_\alpha\\
&+\sum_\alpha\int_{\boldsymbol{b}_\alpha}G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\times\Bigg[\Delta_s h_2(V_s)\\
&\qquad-\int_0^{V_s}h_2(t)\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}\circ\boldsymbol{b}_\alpha)+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}d\Lambda_{s0}(t)\Bigg]d\boldsymbol{b}_\alpha, \qquad (4.28)
\end{aligned}
$$

where $A'(D(t_j;\phi_0))$and $C'(Y_j;D(t_j;\phi_0))$ are the derivatives of $A(D(t_j;\phi))$and $C(Y_j;D(t_j;\phi))$ with respect to ϕ evaluated at ϕ_0 and $B'(\boldsymbol{\beta}_0;\boldsymbol{b})$ is the derivative of $B(\boldsymbol{\beta};\boldsymbol{b})$ with respect to $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta}_0$. Note that (4.28) holds with probability one, so it may not hold for any $V_s\in[0,\tau]$ when $\Delta_s=0$. However, by the similar arguments done in Section 4.4.1, if we integrate both sides from V_s to τ and substract the obtained equation from (4.28) at $\Delta_s=0$ and $V_s=\tau$, it is easily shown that (4.28) also holds for any $V_s\in[0,\tau]$ when $\Delta_s=0$. Hence, the proof of the invertibility of Ω will be completed if we can show $\mathbf{h}_1=0$ and $h_2(t)=0$ from (4.28).

To show $\mathbf{h}_1=0$, particularly we let $\Delta_s=0$ and $V_s=0$ in (4.28) and obtain

$$
\begin{aligned}
0 \;=\;& \sum_\alpha\int_{\boldsymbol{b}_\alpha}G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\times\Bigg[\sum_{j=1}^{n_N}\frac{1}{A(D(t_j;\phi_0))}\left(Y_j\boldsymbol{X}_j-B'(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)\right)\mathbf{h}_1^{\beta}\\
&+\sum_{j=1}^{n_N}\left\{-\left(\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}_0+\widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha)-B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))^2}\right)A'(D(t_j;\phi_0))+C'(Y_j;D(t_j;\phi_0))\right\}\mathrm{h}_1^{\phi}\\
&+\frac{1}{2}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_{\alpha 0})^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_{\alpha 0})-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\right)\\
&+\left(\boldsymbol{b}_\alpha-\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}\right)^T\boldsymbol{\Sigma}_{b0}^{-1}\mathbf{h}_1^{\mu_\alpha}+\frac{1}{w_{\alpha 0}}\mathbf{h}_1^{w_\alpha}\Bigg]d\boldsymbol{b}_\alpha\\
=\;& \sum_\alpha\Bigg[\exp\left\{\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\right\}w_{\alpha 0}\\
&\times\left[\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\left[\exp\left\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\right\}\right]\times\left(\sum_{j=1}^{n_N}\frac{1}{A(D(t_j;\phi_0))}Y_j\boldsymbol{X}_j\mathbf{h}_1^{\beta}\right.\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{j=1}^{n_N}\Big\{-\frac{Y_j\boldsymbol{X}_j\boldsymbol{\beta}_0}{A(D(t_j;\phi_0))^2}A'(D(t_j;\phi_0))+C'(Y_j;D(t_j;\phi_0))\Big\}\ \mathrm{h}_1^{\phi}\\
&+\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{\mu}_{\alpha 0}-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\right)-\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}^T\boldsymbol{\Sigma}_{b0}^{-1}\mathbf{h}_1^{\mu_\alpha}+\frac{1}{w_{\alpha 0}}\mathbf{h}_1^{w_\alpha}\Big)\\
+\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Bigg[&\exp\Big\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\Big\}\times\Big(-\sum_{j=1}^{n_N}\frac{B'(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\mathbf{h}_1^{\beta}\\
&+\sum_{j=1}^{n_N}\Big\{-\Big(\frac{Y_j\widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha-B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))^2}\Big)A'(D(t_j;\phi_0))\Big\}\ \mathrm{h}_1^{\phi}\\
&+\frac{1}{2}\big(\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}_\alpha-2\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{\mu}_\alpha\big)+\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\mathbf{h}_1^{\mu_\alpha}\Big)\Bigg]\Bigg]. \qquad (4.29)
\end{aligned}
$$

We first examine the coefficient for $\boldsymbol{Y}$ in (4.29).

$$
\begin{aligned}
&\sum_{j=1}^{n_N}\Big\{Y_j\Big(\frac{1}{A(D(t_j;\phi_0))}\boldsymbol{X}_j\mathbf{h}_1^{\beta}-\frac{1}{A(D(t_j;\phi_0))^2}\boldsymbol{X}_j\boldsymbol{\beta}_0A'(D(t_j;\phi_0))\ \mathrm{h}_1^{\phi}\Big)\\
&\quad\times\Big[\sum_{\alpha}\Big[\exp\Big\{\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\Big\}w_{\alpha 0}\,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Big[\exp\Big\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\Big\}\Big]\Big]\Big]\Big\}\\
&-\sum_{j=1}^{n_N}\Big\{\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))^2}A'(D(t_j;\phi_0))\ \mathrm{h}_1^{\phi}\\
&\quad\times\Big[\sum_{\alpha}\Big[\exp\Big\{\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\Big\}w_{\alpha 0}\,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Big[\exp\Big\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\Big\}\boldsymbol{b}_\alpha\Big]\Big]\Big]\Big\}\\
&=\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\Big\{\boldsymbol{X}_j\Big(\mathbf{h}_1^{\beta}-\frac{1}{A(D(t_j;\phi_0))}\boldsymbol{\beta}_0A'(D(t_j;\phi_0))\ \mathrm{h}_1^{\phi}\Big)\\
&\quad\times\Big[\sum_{\alpha}\Big[\exp\Big\{\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\Big\}w_{\alpha 0}\,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Big[\exp\Big\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\Big\}\Big]\Big]\Big]\\
&\quad-\frac{\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi_0))}A'(D(t_j;\phi_0))\ \mathrm{h}_1^{\phi}\\
&\quad\times\Big[\sum_{\alpha}\Big[\exp\Big\{\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\Big\}w_{\alpha 0}\,\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Big[\exp\Big\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\Big\}\boldsymbol{b}_\alpha\Big]\Big]\Big]\Big\}\\
&=\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\Big\{\boldsymbol{X}_j\mathbf{h}_1^{\beta}\sum_{\alpha}\Big[\exp\Big\{\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\Big\}w_{\alpha 0}\\
&\qquad\qquad\times\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Big[\exp\Big\{-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\Big\}\Big]\Big]
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{1}{A(D(t_j;\phi_0))} A'(D(t_j;\phi_0))\ \mathrm{h}_1^{\phi}\Bigg(\boldsymbol{X}_j\boldsymbol{\beta}_0 \sum_{\alpha}\Bigg[\exp\Bigg\{ \sum_{j=1}^{n_N} \frac{Y_j}{A(D(t_j;\phi_0))} \widetilde{\boldsymbol{X}}_j \boldsymbol{\mu}_{\alpha 0} \Bigg\} w_{\alpha 0} \\
&\qquad\qquad \times \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Bigg[\exp\Bigg\{ -\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} \Bigg\} \Bigg] \Bigg] \\
&+ \widetilde{\boldsymbol{X}}_j \sum_{\alpha}\Bigg[\exp\Bigg\{ \sum_{j=1}^{n_N} \frac{Y_j}{A(D(t_j;\phi_0))} \widetilde{\boldsymbol{X}}_j \boldsymbol{\mu}_{\alpha 0} \Bigg\} w_{\alpha 0}\, \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Bigg[\exp\Bigg\{ -\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} \Bigg\} \boldsymbol{b}_\alpha \Bigg] \Bigg] \Bigg) \Bigg\} \\
&= 0.
\end{aligned}
$$

Based on assumption (A6), $\mathbf{h}_1^{\beta} = 0$ and $\mathrm{h}_1^{\phi} = 0$.

Then, we examine the constant terms without $\boldsymbol{Y}$ in (4.29). Since $\mathbf{h}_1^{\beta} = 0$ and $\mathrm{h}_1^{\phi} = 0$, (4.29) becomes

$$
\begin{aligned}
&\sum_{\alpha}\Bigg[\exp\Bigg\{ \sum_{j=1}^{n_N} \frac{Y_j}{A(D(t_j;\phi_0))} \widetilde{\boldsymbol{X}}_j \boldsymbol{\mu}_{\alpha 0} \Bigg\} w_{\alpha 0} \\
&\quad \times \Bigg[\mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Bigg[\exp\Bigg\{ -\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} \Bigg\} \Bigg] \times \Bigg(\frac{1}{2} \boldsymbol{\mu}_{\alpha 0}^T \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{\mu}_{\alpha 0} \\
&\qquad\qquad - \frac{1}{2} \mathrm{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \right) - \frac{1}{2} \boldsymbol{\mu}_{\alpha 0}^T \boldsymbol{\Sigma}_{b0}^{-1} \mathbf{h}_1^{\mu_\alpha} + \frac{1}{w_{\alpha 0}} \mathbf{h}_1^{w_\alpha} \Bigg) \\
&\quad + \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha}\Bigg[\exp\Bigg\{ -\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} \Bigg\} \times \Bigg(\frac{1}{2} \boldsymbol{b}_\alpha^T \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{b}_\alpha \\
&\qquad\qquad - \boldsymbol{b}_\alpha^T \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{\mu}_\alpha + \boldsymbol{b}_\alpha^T \boldsymbol{\Sigma}_{b0}^{-1} \mathbf{h}_1^{\mu_\alpha} \Bigg) \Bigg] \Bigg] \Bigg] \\
&= \mathrm{E}_{\alpha,\boldsymbol{b}}\Bigg[\exp\Bigg\{ \sum_{j=1}^{n_N} \frac{Y_j}{A(D(t_j;\phi_0))} \widetilde{\boldsymbol{X}}_j \boldsymbol{\mu}_{\alpha 0} \Bigg\} \times \exp\Bigg\{ -\sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} \Bigg\} \\
&\quad \times \Bigg(\frac{1}{2} \boldsymbol{\mu}_{\alpha 0}^T \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{\mu}_{\alpha 0} - \frac{1}{2} \mathrm{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \right) \\
&\qquad - \frac{1}{2} \boldsymbol{\mu}_{\alpha 0}^T \boldsymbol{\Sigma}_{b0}^{-1} \mathbf{h}_1^{\mu_\alpha} + \frac{1}{w_{\alpha 0}} \mathbf{h}_1^{w_\alpha} + \frac{1}{2} \boldsymbol{b}_\alpha^T \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{b}_\alpha - \boldsymbol{b}_\alpha^T \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{D}_b \boldsymbol{\Sigma}_{b0}^{-1} \boldsymbol{\mu}_\alpha + \boldsymbol{b}_\alpha^T \boldsymbol{\Sigma}_{b0}^{-1} \mathbf{h}_1^{\mu_\alpha} \Bigg) \Bigg] \\
&= 0,
\end{aligned}
$$

where $\boldsymbol{b}$ follows a multivariate normal distribution with mean $\boldsymbol{\Sigma}_{b0}\big(\sum_{j=1}^{n_N} (Y_j \widetilde{\boldsymbol{X}}_j / A(D(t_j;\phi_0))) \big) + \boldsymbol{\mu}_{\alpha 0}$ and covariance $\boldsymbol{\Sigma}_{b0}$. For any fixed $\widetilde{\boldsymbol{X}}$, treating $\boldsymbol{X}^T\boldsymbol{Y}$ as a parameter

in this normal family, $\boldsymbol{b} = \sum_{k=1}^{K} I(\alpha = k)\boldsymbol{b}_k$ is the complete statistic for $\boldsymbol{X}^T\boldsymbol{Y}$, therefore,

$$\exp\left\{\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\right\}\times\left(\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{\mu}_{\alpha 0}-\frac{1}{2}\mathrm{Tr}\big(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\big)\right.$$
$$\left.-\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}^T\boldsymbol{\Sigma}_{b0}^{-1}\mathbf{h}_1^{\mu_\alpha}+\frac{1}{w_{\alpha 0}}\mathbf{h}_1^{w_\alpha}+\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}_\alpha-\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{\mu}_\alpha+\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\mathbf{h}_1^{\mu_\alpha}\right)$$
$$= 0.$$

Since $\exp\big\{\sum_{j=1}^{n_N}\frac{Y_j}{A(D(t_j;\phi_0))}\widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}-\sum_{j=1}^{n_N}\frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))}\big\}\neq 0$, by (A6), we have

$$\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{\mu}_{\alpha 0}-\frac{1}{2}\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\right)+\left(-\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}^T+\boldsymbol{b}_\alpha^T\right)\boldsymbol{\Sigma}_{b0}^{-1}\mathbf{h}_1^{\mu_\alpha}+\frac{1}{w_{\alpha 0}}\mathbf{h}_1^{w_\alpha}$$
$$+\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{b}_\alpha-\boldsymbol{b}_\alpha^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{\mu}_\alpha \;=\; 0.$$

$\because -\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}^T+\boldsymbol{b}_\alpha^T\neq 0$ and $\boldsymbol{\Sigma}_{b0}^{-1}\neq 0 \Longrightarrow \therefore \mathbf{h}_1^{\mu_\alpha}=0$, $\alpha=1,\ldots,K$, by (A6).

$\because 1/w_\alpha\neq 0 \Longrightarrow \therefore \mathbf{h}_1^{w_\alpha}=0$, $\alpha=1,\ldots,K$, by (A6).

$\because \boldsymbol{\Sigma}_{b0}^{-1}\neq 0 \Longrightarrow \therefore \boldsymbol{D}_b=\mathbf{0}$ by (A6).

Next, we let $\Delta_s=0$ in (4.28) and obtain

$$0=\sum_\alpha\int_{\boldsymbol{b}_\alpha}G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\times\left[\sum_{j=1}^{n_N}\frac{1}{A(D(t_j;\phi_0))}\big(Y_j\boldsymbol{X}_j-B'(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)\big)\mathbf{h}_1^\beta\right.$$
$$+\sum_{j=1}^{n_N}\left\{-\left(\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}_0+\widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha)-B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))^2}\right)A'(D(t_j;\phi_0))+C'(Y_j;D(t_j;\phi_0))\right\}\,\mathrm{h}_1^\phi$$
$$+\frac{1}{2}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_{\alpha 0})^T\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_{b0}^{-1}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_{\alpha 0})-\frac{1}{2}\,\mathrm{Tr}\left(\boldsymbol{\Sigma}_{b0}^{-1}\boldsymbol{D}_b\right)$$
$$+\left(\boldsymbol{b}_\alpha-\frac{1}{2}\boldsymbol{\mu}_{\alpha 0}\right)^T\boldsymbol{\Sigma}_{b0}^{-1}\mathbf{h}_1^{\mu_\alpha}+\frac{1}{w_{\alpha 0}}\mathbf{h}_1^{w_{\alpha 0}}$$
$$\left.-\int_0^{V_s}\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0\circ\boldsymbol{b}_\alpha)+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}\times\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{b}_\alpha^T)\mathbf{h}_1^\psi+\boldsymbol{Z}(t)\mathbf{h}_1^\gamma\right\}d\Lambda_{s0}(t)\right]d\boldsymbol{b}_\alpha$$
$$+\sum_\alpha\int_{\boldsymbol{b}_\alpha}G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta}_0,\Lambda_{s0})\times\left[-\int_0^{V_s}h_2(t)\exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}\circ\boldsymbol{b}_\alpha)+\boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\}d\Lambda_{s0}(t)\right]d\boldsymbol{b}_\alpha.$$

Since $\mathbf{h}_1^\beta=0$, h_1^ϕ, $\mathbf{h}_1^{\mu_\alpha}=0$, $\mathbf{h}_1^{w_{\alpha 0}}=0$, $\alpha=1,\ldots,K$, and $\boldsymbol{D}_b=0$, the above expression can

be written as

$$0 = \mathrm{E}_{\alpha,\boldsymbol{b}}\Bigg[\exp\left\{ \sum_{j=1}^{n_N} \frac{Y_j}{A(D(t_j;\phi_0))}\big(\boldsymbol{X}_j\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\big) - \sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} \right\} \\ \times \int_0^{V_s} \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0 \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\} \times \left[(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{b}_\alpha^T)\mathbf{h}_1^\psi + \boldsymbol{Z}(t)\mathbf{h}_1^\gamma + h_2(t)\right] d\Lambda_{s0}(t) \Bigg], \tag{4.30}$$

where $\boldsymbol{b}_\alpha$ follows a multivariate normal distribution with mean $\boldsymbol{\Sigma}_{b0}\big[\sum_{j=1}^{n_N}\big(Y_j\widetilde{\boldsymbol{Z}}_j/A(D(t_j;\phi_0))\big)\big] + \boldsymbol{\mu}_{\alpha 0}$ and covariance $\boldsymbol{\Sigma}_{b0}$. Likewise, for any fixed $\widetilde{\boldsymbol{X}}$, treating $\boldsymbol{X}^T\boldsymbol{Y}$ as a parameter in this normal family, $\boldsymbol{b}_\alpha$ is the complete statistic for $\boldsymbol{X}^T\boldsymbol{Y}$, therefore,

$$\exp\left\{ \sum_{j=1}^{n_N} \frac{Y_j}{A(D(t_j;\phi_0))}\big(\boldsymbol{X}_j\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\big) - \sum_{j=1}^{n_N} \frac{B(\boldsymbol{\beta}_0;\boldsymbol{b}_\alpha)}{A(D(t_j;\phi_0))} \right\} \\ \times \int_0^{V_s} \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0 \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\} \times \left[(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{b}_\alpha^T)\mathbf{h}_1^\psi + \boldsymbol{Z}(t)\mathbf{h}_1^\gamma + h_2(t)\right] d\Lambda_{s0}(t) \\ = 0.$$

Since $\exp\big\{\sum_{j=1}^{n_N}\big[Y_j\big(\boldsymbol{X}_j\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{X}}_j\boldsymbol{\mu}_{\alpha 0}\big)/A(D(t_j;\phi_0))\big] - \sum_{j=1}^{n_N}\big[B(\boldsymbol{\beta}_0;\boldsymbol{b})/A(D(t_j;\phi_0))\big]\big\} \neq 0$, equivalently

$$\int_0^{V_s} \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}_0 \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}_0\right\} \times \left[(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{b}_\alpha^T)\mathbf{h}_1^\psi + \boldsymbol{Z}(t)\mathbf{h}_1^\gamma + h_2(t)\right] d\Lambda_{s0}(t) = 0$$

by assumption (A6). From assumption (A6), this immediately gives $\mathbf{h}_1^\psi = 0$, $\mathbf{h}_1^\gamma = 0$ and $h_2(t) = 0$. Hence, the proof of condition (c) is completed.

Since the conditions (a)–(d) have been proved, Theorem 3.3.1 of van der Vaart and Wellner (1996) concludes that $\sqrt{m_s}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda}_s - \Lambda_{s0})$ weakly converges to a tight random element in $\ell^\infty(\mathscr{H})$. Furthermore, we obtain

$$\sqrt{m_s}\nabla U_{\xi_0}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}_s-\Lambda_{s0})[\mathbf{h}_1,h_2]$$
$$=\sqrt{m_s}(\mathbf{P}_{m_s}-\mathbf{P})\big\{l_\theta(\boldsymbol{\theta}_0,\Lambda_{s0})^T\mathbf{h}_1+l_{\Lambda_s}(\boldsymbol{\theta}_0,\Lambda_{s0})[h_2]\big\}+o_P(1), \tag{4.31}$$

where $o_P(1)$ is a random variable which converges to zero in probability in $\ell^\infty(\mathscr{H})$. On the other hand, from (4.27), we have

$$\sqrt{m_s}\nabla U_{\xi_0}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}_s-\Lambda_{s0})[\mathbf{h}_1,h_2]$$
$$=\sqrt{m_s}\big\{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\Omega_1[\mathbf{h}_1,h_2]+\int_0^\tau\Omega_2[\mathbf{h}_1,h_2]d(\widehat{\Lambda}_s-\Lambda_{s0})(t)\big\}. \tag{4.32}$$

By denoting $(\mathbf{h}_1^*,h_2^*)=\Omega^{-1}(\mathbf{h}_1,h_2)$, we have $(\mathbf{h}_1,h_2)=\Omega(\mathbf{h}_1^*,h_2^*)$, and replacing $(\mathbf{h}_1,h_2)$ with $(\mathbf{h}_1^*,h_2^*)$ in (4.31) and (4.32) leads to the followings, respectively.

$$\sqrt{m_s}\nabla U_{\xi_0}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}_s-\Lambda_{s0})[\mathbf{h}_1^*,h_2^*]$$
$$=\sqrt{m_s}(\mathbf{P}_{m_s}-\mathbf{P})\big\{l_\theta(\boldsymbol{\theta}_0,\Lambda_{s0})^T\mathbf{h}_1^*+l_{\Lambda_s}(\boldsymbol{\theta}_0,\Lambda_{s0})[h_2^*]\big\}+o_P(1),$$

and

$$\sqrt{m_s}\nabla U_{\xi_0}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0,\widehat{\Lambda}_s-\Lambda_{s0})[\mathbf{h}_1^*,h_2^*]$$
$$=\sqrt{m_s}\left\{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\Omega_1[\mathbf{h}_1^*,h_2^*]+\int_0^\tau\Omega_2[\mathbf{h}_1^*,h_2^*]d(\widehat{\Lambda}_s-\Lambda_{s0})(t)\right\}$$
$$=\sqrt{m_s}\left\{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\mathbf{h}_1+\int_0^\tau h_2(t)d(\widehat{\Lambda}_s-\Lambda_{s0})(t)\right\}.$$

Thus, we obtain

$$\sqrt{m_s}\left\{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\mathbf{h}_1+\int_0^\tau h_2(t)d(\widehat{\Lambda}_s-\Lambda_{s0})(t)\right\}$$
$$=\sqrt{m_s}(\mathbf{P}_{m_s}-\mathbf{P})\big\{l_\theta(\boldsymbol{\theta}_0,\Lambda_{s0})^T\mathbf{h}_1^*+l_{\Lambda_s}(\boldsymbol{\theta}_0,\Lambda_{s0})[h_2^*]\big\}+o_P(1). \tag{4.33}$$

Note that the first term on the right-hand side in (4.33) is $\sqrt{m_s}\{U_{m_s}(\boldsymbol{\theta}_0, \Lambda_{s0}) - U_s(\boldsymbol{\theta}_0, \Lambda_{s0})\}$, which is an empirical process in the space $\ell^\infty(\mathscr{H})$, and it is shown that $\mathscr{G}$ is P-Donsker in Section 4.4.3.2. Therefore, $\sqrt{m_s}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \widehat{\Lambda}_s - \Lambda_{s0})$ weakly converges to a Gaussian process in $\ell^\infty(\mathscr{H})$.

In particular, if we choose $h_2 = 0$ in (4.33), then $\widehat{\boldsymbol{\theta}}^T \mathbf{h}_1$ is an asymptotic linear estimator for $\boldsymbol{\theta}_0^T \mathbf{h}_1$ with influence function being $l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T \mathbf{h}_1^* + l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2^*]$. Since this influence function is in the linear space spanned by the score functions for $\boldsymbol{\theta}_0$ and Λ_{s0}, Proposition 3.3.1 (p65) in Bickel, Klaassen, Ritov and Wellner (1993) concludes that the influence function is the same as the efficient influence function for $\boldsymbol{\theta}_0^T \mathbf{h}_1$; that is $\widehat{\boldsymbol{\theta}}$ is an efficient estimator for $\boldsymbol{\theta}_0$ and Theorem 4.2 is proved.

4.4.3 Supplementary proofs

The proofs for P-Donsker property of the classes $\mathscr{F}$ and $\mathscr{G}$ needed in Sections 4.4.1 and 4.4.2 are presented in Sections 4.4.3.1~4.4.3.2 respectively. In Section 4.4.3.3, we prove *Fréchet* differentiability of $U(\xi)$ at ξ_0 and derive the derivative operator ∇U_{ξ_0} used in Section 4.4.2.

4.4.3.1 Proof of P-Donsker property of $\mathscr{F}$

We defined that a class $\mathscr{F} = \{Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) : v \in [0, \tau], \boldsymbol{\theta} \in \Theta, \Lambda_s \in \mathscr{A}, s = 1, \ldots, S\}$, where $\mathscr{A} = \{\Lambda_s \in \mathbb{W}, \Lambda_s(0) = 0, \Lambda_s(\tau) \leq B_{s0}, s = 1, \ldots, S\}$, B_{s0} is the constant given in the second step and $\mathbb{W}$ contains all nondecreasing functions in $[0, \tau]$. We can rewrite $Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)$ as

$$Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) = Q_1(v, \boldsymbol{O}; \boldsymbol{\theta}) \frac{Q_2(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)}{Q_3(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)},$$

where

$$
\begin{aligned}
&Q_1(v, \boldsymbol{O}; \boldsymbol{\theta}) \\
&= \exp\left\{\boldsymbol{Z}(v)\boldsymbol{\gamma} + \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)\right)\Sigma_b(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)^T + \frac{1}{2}R(v)\right\}, \\
&Q_2(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) \\
&= \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} \exp\left\{ -\frac{1}{2}\boldsymbol{b}_\alpha^T \boldsymbol{b}_\alpha + \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)\right)\boldsymbol{\mu}_\alpha - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta}; \boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \right. \\
&\quad - \int_0^{V_s} \exp\left\{ (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}_\alpha + \boldsymbol{Z}(t)\boldsymbol{\gamma} \right. \\
&\qquad \left. \left. + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\left[\boldsymbol{\Sigma}_b\left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\right)^T + \boldsymbol{\mu}_\alpha\right] + R(t)\right\} d\Lambda_s(t)\right\} w_\alpha d\boldsymbol{b}_\alpha, \\
&Q_3(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) \\
&= \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} \exp\left\{ -\frac{1}{2}\boldsymbol{b}_\alpha^T \boldsymbol{b}_\alpha + \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)\right)\boldsymbol{\mu}_\alpha - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta}; \boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \right. \\
&\quad - \int_0^{V_s} \exp\left\{ (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}_\alpha + \boldsymbol{Z}(t)\boldsymbol{\gamma} \right. \\
&\qquad \left. \left. + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\left[\boldsymbol{\Sigma}_b\left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\right)^T + \boldsymbol{\mu}_\alpha\right]\right\} d\Lambda_s(t)\right\} w_\alpha d\boldsymbol{b}_\alpha,
\end{aligned}
$$

$R(t) = (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)\Sigma_b(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T)^T$, $R(v)$ is $R(t)$ evaluated at $t = v$, $B_1(\boldsymbol{\beta}; \boldsymbol{b}_\alpha) = B(\boldsymbol{\beta}; g_1(\boldsymbol{b}_\alpha))$, $B_2(\boldsymbol{\beta}; \boldsymbol{b}_\alpha) = B(\boldsymbol{\beta}; g_2(\boldsymbol{b}_\alpha))$, $g_1(\boldsymbol{b}_\alpha) = \boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}_\alpha + \boldsymbol{\Sigma}_b\big[\sum_{j=1}^{n_N}\big(Y_j\widetilde{\boldsymbol{X}}_j/A(D(t_j;\phi))\big) + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)\big]^T + \boldsymbol{\mu}_\alpha$ and $g_2(\boldsymbol{b}_\alpha) = \boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}_\alpha + \boldsymbol{\Sigma}_b\big[\sum_{j=1}^{n_N}\big(Y_j\widetilde{\boldsymbol{X}}_j/A(D(t_j;\phi))\big) + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)\big]^T + \boldsymbol{\mu}_\alpha$.

Using assumption (A3), we can easily show that $Q_1(v, \boldsymbol{O}; \boldsymbol{\theta})$ is continuously differentiable with respect to v and $\boldsymbol{\theta}$, and

$$
\|\nabla_{\boldsymbol{\theta}} Q_1(v, \boldsymbol{O}; \boldsymbol{\theta})\| + \left|\frac{d}{dv} Q_1(v, \boldsymbol{O}; \boldsymbol{\theta})\right| \;\le\; e^{k_1 + k_2\|\boldsymbol{Y}\|}
$$

for some positive constants k_1 and k_2. Furthermore, it holds that

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}} Q_2(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)\| + \left| \frac{d}{dv} Q_2(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) \right| \\
&\quad \leq \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} \left[\exp\left\{ -\frac{1}{2}\boldsymbol{b}_\alpha^T \boldsymbol{b}_\alpha + \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \right) \boldsymbol{\mu}_\alpha \right. \right. \\
&\qquad\qquad \left. \left. - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta}; \boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \right\} \times e^{k_3 \|\boldsymbol{b}_\alpha\| + k_4 \|\boldsymbol{Y}\|_{+k_5(\alpha)}} \times B_{s0} \times w_\alpha \right] d\boldsymbol{b}_\alpha \\
&\quad \leq e^{k_6 + k_7 \|\boldsymbol{Y}\|}
\end{aligned}
$$

and $\quad \|\nabla_{\boldsymbol{\theta}} Q_3(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s)\| + \left| \frac{d}{dv} Q_3(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) \right| \leq e^{k_8 + k_9 \|\boldsymbol{Y}\|}$

for some positive constants k_3, k_4, k_6, k_7, k_8, and k_9, and a deterministic function of α, $k(\alpha)$. Additionally, note that, for any $0 < \Lambda < \infty$, $0 < e^{-\Lambda} < 1$ and $e^{-\Lambda} < \Lambda$ and thus $e^{-\Lambda_1} - e^{-\Lambda_2} < \Lambda_1 - \Lambda_2$ for any Λ_1 and Λ_2 over $(0, \infty)$. Hence,

$$
\begin{aligned}
&|Q_2(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_{s1}) - Q_2(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_{s2})| \\
&= \left| \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} \exp\left\{ -\frac{1}{2}\boldsymbol{b}_\alpha^T \boldsymbol{b}_\alpha + \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \right) \boldsymbol{\mu}_\alpha \right. \right. \\
&\quad - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta}; \boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \left[\exp\left\{ -\int_0^v \exp\left\{ (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \boldsymbol{b}_\alpha + \boldsymbol{Z}(t)\boldsymbol{\gamma} \right. \right. \right. \\
&\qquad + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \left[\boldsymbol{\Sigma}_b \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \right)^T + \boldsymbol{\mu}_\alpha \right] \\
&\qquad \left. \left. + R(t) \right\} d\Lambda_{s1}(t) \right\} - \exp\left\{ -\int_0^v \exp\left\{ (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \boldsymbol{b}_\alpha + \boldsymbol{Z}(t)\boldsymbol{\gamma} \right. \right. \\
&\qquad + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \left[\boldsymbol{\Sigma}_b \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \right)^T + \boldsymbol{\mu}_\alpha \right] \\
&\qquad \left. \left. \left. + R(t) \right\} d\Lambda_{s2}(t) \right\} \right] w_\alpha d\boldsymbol{b}_\alpha \Bigg| \\
&\leq \left| \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} \exp\left\{ -\frac{1}{2}\boldsymbol{b}_\alpha^T \boldsymbol{b}_\alpha + \left(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \right) \boldsymbol{\mu}_\alpha \right. \right. \\
&\quad \left. \left. - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta}; \boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \right\} \left[\int_0^v \exp\left\{ (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \boldsymbol{b}_\alpha + \boldsymbol{Z}(t)\boldsymbol{\gamma} \right. \right. \right.
\end{aligned}
$$

$$
\begin{aligned}
&+ (\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg[\boldsymbol{\Sigma}_b\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+\boldsymbol{\mu}_\alpha\bigg]\\
&+ R(t)\bigg\}d\big(\Lambda_{s1}-\Lambda_{s2}\big)(t)\bigg]w_\alpha d\boldsymbol{b}_\alpha\bigg|\\
=&\bigg|\sum_\alpha\int_{\boldsymbol{b}_\alpha}\exp\bigg\{\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+(\Delta+1)(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\bigg)\boldsymbol{\mu}_\alpha-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))}\bigg\}\times(2\pi)^{d_b/2}\\
&\times(2\pi)^{-d_b/2}\times\bigg[\int_0^v\exp\bigg\{-\frac{1}{2}\big[\boldsymbol{b}_\alpha-\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T\big]^T\big[\boldsymbol{b}_\alpha-\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T\big]\bigg\}\\
&\times\exp\bigg\{\frac{1}{2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T+\boldsymbol{Z}(t)\boldsymbol{\gamma}\\
&+ (\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg[\boldsymbol{\Sigma}_b\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+\boldsymbol{\mu}_\alpha\bigg]\\
&+ R(t)\bigg\}d\big(\Lambda_{s1}-\Lambda_{s2}\big)(t)\bigg]w_\alpha d\boldsymbol{b}_\alpha\bigg|\\
\le&\bigg|\sum_\alpha\bigg[\exp\bigg\{\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+(\Delta+1)(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\bigg)\boldsymbol{\mu}_\alpha\bigg\}\times(2\pi)^{d_b/2}\\
&\times\int_0^v\bigg[\exp\bigg\{\frac{1}{2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T+\boldsymbol{Z}(t)\boldsymbol{\gamma}\\
&+ (\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg[\boldsymbol{\Sigma}_b\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+\boldsymbol{\mu}_\alpha\bigg]+R(t)\bigg\}\\
&\times(2\pi)^{-d_b/2}\int_{\boldsymbol{b}_\alpha}\exp\bigg\{-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))}\bigg\}\\
&\times\exp\bigg\{-\frac{1}{2}\big[\boldsymbol{b}_\alpha-\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T\big]^T\big[\boldsymbol{b}_\alpha-\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T\big]\bigg\}d\boldsymbol{b}_\alpha\bigg]\\
&d\big(\Lambda_{s1}-\Lambda_{s2}\big)(t)\bigg]w_\alpha\bigg|\\
=&\bigg|\sum_\alpha\bigg[\exp\bigg\{\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+(\Delta+1)(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\bigg)\boldsymbol{\mu}_\alpha\bigg\}\times(2\pi)^{d_b/2}\\
&\times\int_0^v\bigg[\exp\bigg\{\frac{1}{2}\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)\big((\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\big)^T+\boldsymbol{Z}(t)\boldsymbol{\gamma}\\
&+ (\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg[\boldsymbol{\Sigma}_b\bigg(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\bigg)^T+\boldsymbol{\mu}_\alpha\bigg]+R(t)\bigg\}\\
&\times\mathrm{E}_{\boldsymbol{b}_{\alpha|\alpha}}\bigg[\exp\bigg\{-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))}\bigg\}\bigg]\bigg]d\big(\Lambda_{s1}-\Lambda_{s2}\big)(t)\bigg]w_\alpha\bigg|
\end{aligned}
$$

$$
\begin{aligned}
&= \Bigg| \sum_{\alpha} \Bigg[\exp\Bigg\{ \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg) \boldsymbol{\mu}_\alpha \Bigg\} \times (2\pi)^{d_b/2} \\
&\quad \times \Bigg[\big(\Lambda_{s1}(v) - \Lambda_{s2}(v)\big) \exp\Bigg\{ \frac{1}{2} \big((\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big) \big((\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big)^T + \boldsymbol{Z}(t)\boldsymbol{\gamma} \\
&\qquad + (\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg[\boldsymbol{\Sigma}_b \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg)^T + \boldsymbol{\mu}_\alpha \Bigg] + R(v) \Bigg\} \\
&\qquad\qquad \times \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha} \Bigg[\exp\Bigg\{ - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \Bigg\} \Bigg] \\
&\quad - \int_0^v \Bigg[\big(\Lambda_{s1}(t) - \Lambda_{s2}(t)\big) \frac{d}{dt} \Bigg[\exp\Bigg\{ \frac{1}{2} \big((\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big) \big((\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big)^T + \boldsymbol{Z}(t)\boldsymbol{\gamma} \\
&\qquad + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \Bigg[\boldsymbol{\Sigma}_b \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \Bigg)^T + \boldsymbol{\mu}_\alpha \Bigg] + R(t) \Bigg\} \\
&\qquad \times \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha} \Bigg[\exp\Bigg\{ - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \Bigg\} \Bigg] \Bigg] \Bigg] dt \Bigg] \Bigg] w_\alpha \Bigg| \\
&\leq \sum_{\alpha} w_\alpha \Bigg[\exp\Bigg\{ \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg) \boldsymbol{\mu}_\alpha \Bigg\} \times (2\pi)^{d_b/2} \\
&\quad \times \Bigg[\big|\Lambda_{s1}(v) - \Lambda_{s2}(v)\big| \exp\Bigg\{ \frac{1}{2} \big((\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big) \big((\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big)^T + \boldsymbol{Z}(t)\boldsymbol{\gamma} \\
&\qquad + (\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg[\boldsymbol{\Sigma}_b \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg)^T + \boldsymbol{\mu}_\alpha \Bigg] + R(v) \Bigg\} \\
&\qquad\qquad \times \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha} \Bigg[\exp\Bigg\{ - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \Bigg\} \Bigg] \\
&\quad + \int_0^v \Bigg[\big|\Lambda_{s1}(t) - \Lambda_{s2}(t)\big| \Bigg| \frac{d}{dt} \Bigg[\exp\Bigg\{ \frac{1}{2} \big((\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big) \big((\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big)^T + \boldsymbol{Z}(t)\boldsymbol{\gamma} \\
&\qquad + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \Bigg[\boldsymbol{\Sigma}_b \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \Bigg)^T + \boldsymbol{\mu}_\alpha \Bigg] + R(t) \Bigg\} \\
&\qquad \times \mathrm{E}_{\boldsymbol{b}_\alpha|\alpha} \Bigg[\exp\Bigg\{ - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \Bigg\} \Bigg] \Bigg| \Bigg] dt \Bigg] \\
&= \big|\Lambda_{s1}(v) - \Lambda_{s2}(v)\big| \exp\Bigg\{ \frac{1}{2} \big((\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big) \big((\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big)^T + \boldsymbol{Z}(t)\boldsymbol{\gamma} \\
&\qquad + (\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg)^T + R(v) \Bigg\} \times (2\pi)^{d_b/2} \\
&\qquad \times \mathrm{E}_{\alpha,\boldsymbol{b}} \Bigg[\exp\Bigg\{ \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg) \boldsymbol{\mu}_\alpha - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \Bigg\} \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+ \int_0^v \Bigg[\big|\Lambda_{s1}(t) - \Lambda_{s2}(t)\big| \times (2\pi)^{d_b/2} \\
&\quad \times \mathrm{E}_{\alpha,\boldsymbol{b}} \Bigg[\exp\Bigg\{ \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg) \boldsymbol{\mu}_\alpha \Bigg\} \\
&\quad \times \Bigg| \frac{d}{dt} \Bigg[\exp\Bigg\{ \frac{1}{2} \big((\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big) \big((\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big)^T + \boldsymbol{Z}(t)\boldsymbol{\gamma} \\
&\quad + (\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \Bigg[\boldsymbol{\Sigma}_b \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(t) \circ \boldsymbol{\psi}^T) \Bigg)^T + \boldsymbol{\mu}_\alpha \Bigg] \\
&\quad + R(t) - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \Bigg\} \Bigg] \Bigg| \Bigg] dt \\
&\leq (2\pi)^{d_b/2} \exp\Bigg\{ \frac{1}{2} \big((\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big) \big((\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b^{1/2} \big)^T + \boldsymbol{Z}(t)\boldsymbol{\gamma} \\
&\quad + (\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \boldsymbol{\Sigma}_b \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg)^T + R(v) \Bigg\} \\
&\quad \times \mathrm{E}_{\alpha,\boldsymbol{b}} \Bigg[\exp\Bigg\{ \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + (\Delta+1)(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg) \boldsymbol{\mu}_\alpha - \sum_{j=1}^{n_N} \frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \Bigg\} \Bigg] \\
&\quad \times \Bigg[\big|\Lambda_{s1}(v) - \Lambda_{s2}(v)\big| + \int_0^v \big|\Lambda_{s1}(t) - \Lambda_{s2}(t)\big| dt \Bigg] \\
&\leq e^{k_{10}+k_{11}\|\boldsymbol{Y}\|} \Bigg[\big|\Lambda_{s1}(v) - \Lambda_{s2}(v)\big| + \int_0^\tau \big|\Lambda_{s1}(t) - \Lambda_{s2}(t)\big| dt \Bigg],
\end{aligned}
$$

where k_{10} and k_{11} are positive constants. Similarly,

$$
\begin{aligned}
&|Q_3(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_{s1}) - Q_3(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_{s2})| \\
&\quad \leq \; e^{k_{12}+k_{13}\|\boldsymbol{Y}\|} \Bigg[\big|\Lambda_{s1}(v) - \Lambda_{s2}(v)\big| + \int_0^\tau \big|\Lambda_{s1}(t) - \Lambda_{s2}(t)\big| dt \Bigg],
\end{aligned}
$$

where k_{12} and k_{13} are positive constants.

On the other hand, there exist positive constants $k_{14}, \ldots, k_{26}$ such that

$$
\begin{aligned}
&\big|Q_1(v,\boldsymbol{O};\boldsymbol{\theta})\big| \\
&= \Bigg| \exp\Bigg\{ \boldsymbol{Z}(v)\boldsymbol{\gamma} + \Bigg(\sum_{j=1}^{n_N} \frac{Y_j \widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T) \Bigg) \Sigma_b (\widetilde{\boldsymbol{Z}}(v) \circ \boldsymbol{\psi}^T)^T + \frac{1}{2} R(v) \Bigg\} \Bigg|
\end{aligned}
$$

$$
\begin{aligned}
&\leq e^{k_{14}+k_{15}\|\boldsymbol{Y}\|},\\
&\left|Q_2(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\right|\\
&=\left|\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{-\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{b}_\alpha+\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+(\Delta+1)(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\right)\boldsymbol{\mu}_\alpha\right.\\
&\quad-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))}-\int_0^{V_s}\exp\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}_\alpha+\boldsymbol{Z}(t)\boldsymbol{\gamma}\right.\\
&\quad+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\left[\boldsymbol{\Sigma}_b\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\right)^T+\boldsymbol{\mu}_\alpha\right]\\
&\quad\left.\left.+R(t)\right\}d\Lambda_s(t)\right\}w_\alpha d\boldsymbol{b}_\alpha\Bigg|\\
&\leq\left|\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{-\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{b}_\alpha+\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+(\Delta+1)(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\right)\boldsymbol{\mu}_\alpha\right.\right.\\
&\quad\left.-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))}\right\}\times\left[2\int_0^{v}\exp\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}_\alpha+\boldsymbol{Z}(t)\boldsymbol{\gamma}\right.\right.\\
&\quad+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\left[\boldsymbol{\Sigma}_b\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\right)^T+\boldsymbol{\mu}_\alpha\right]\\
&\quad\left.\left.\left.+R(t)\right\}d\Lambda_s(t)\right]w_\alpha d\boldsymbol{b}_\alpha\Bigg|\\
&\leq\left|\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{-\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{b}_\alpha+\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+(\Delta+1)(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\right)\boldsymbol{\mu}_\alpha\right.\right.\\
&\quad\left.-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))}\right\}\times 2\times\exp\left\{k_{16}\|\boldsymbol{b}_\alpha)\|+k_{17}\|\boldsymbol{Y}\|+k_{18}\|\boldsymbol{\mu}_\alpha\|+k_{19}\right\}\times B_{s0}\times w_\alpha d\boldsymbol{b}_\alpha\Bigg|\\
&\leq e^{k_{19}+k_{20}\|\boldsymbol{Y}\|},
\end{aligned}
$$

and

$$
\begin{aligned}
&Q_3(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\\
&=\sum_{\alpha}\int_{\boldsymbol{b}_\alpha}\exp\left\{-\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{b}_\alpha+\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\right)\boldsymbol{\mu}_\alpha-\sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))}\right.\\
&\quad-\int_0^{V_s}\exp\left\{(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\boldsymbol{\Sigma}_b^{1/2}\boldsymbol{b}_\alpha+\boldsymbol{Z}(t)\boldsymbol{\gamma}\right.\\
&\quad\left.\left.+(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\left[\boldsymbol{\Sigma}_b\left(\sum_{j=1}^{n_N}\frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))}+\Delta(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{\psi}^T)\right)^T+\boldsymbol{\mu}_\alpha\right]\right\}d\Lambda_s(t)\right\}w_\alpha d\boldsymbol{b}_\alpha
\end{aligned}
$$

$$
\begin{aligned}
&\geq \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} \exp\Bigg\{ -\frac{1}{2}\boldsymbol{b}_\alpha^T\boldsymbol{b}_\alpha + \Bigg(\sum_{j=1}^{n_N} \frac{Y_j\widetilde{\boldsymbol{X}}_j}{A(D(t_j;\phi))} + \Delta(\widetilde{\boldsymbol{Z}}(v)\circ\boldsymbol{\psi}^T)\Bigg)\boldsymbol{\mu}_\alpha - \sum_{j=1}^{n_N}\frac{B_1(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))} \\
&\quad - \exp\left\{k_{22}\|\boldsymbol{b}_\alpha\| + k_{23}\|\boldsymbol{Y}\| + k_{24}\|\boldsymbol{\mu}_\alpha\| + k_{25}\right\} \times B_{s0}\Bigg\} w_\alpha d\boldsymbol{b}_\alpha \\
&\geq k_{26} > 0.
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}} Q(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\| + \left|\frac{d}{dv}Q(v,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)\right| \\
&= \left\|(\nabla_{\boldsymbol{\theta}}Q_1)\frac{Q_2}{Q_3} + Q_1\left(\nabla_{\boldsymbol{\theta}}\frac{Q_2}{Q_3}\right)\right\| + \left|\left(\frac{d}{dv}Q_1\right)\frac{Q_2}{Q_3} + Q_1\left(\frac{d}{dv}\left(\frac{Q_2}{Q_3}\right)\right)\right| \\
&= \left\|(\nabla_{\boldsymbol{\theta}}Q_1)\frac{Q_2}{Q_3} + Q_1\left[(\nabla_{\boldsymbol{\theta}}Q_2)\frac{1}{Q_3} + Q_2\frac{(-1)}{Q_3^2}(\nabla_{\boldsymbol{\theta}}Q_3)\right]\right\| \\
&\qquad + \left|\left(\frac{d}{dv}Q_1\right)\frac{Q_2}{Q_3} + Q_1\left[\left(\frac{d}{dv}Q_2\right)\frac{1}{Q_3} + Q_2\frac{(-1)}{Q_3^2}\left(\frac{d}{dv}Q_3\right)\right]\right| \\
&= \left\|(\nabla_{\boldsymbol{\theta}}Q_1)\frac{Q_2}{Q_3} + (\nabla_{\boldsymbol{\theta}}Q_2)\frac{Q_1}{Q_3} - (\nabla_{\boldsymbol{\theta}}Q_3)\frac{Q_1Q_2}{Q_3^2}\right\| \\
&\qquad + \left|\left(\frac{d}{dv}Q_1\right)\frac{Q_2}{Q_3} + \left(\frac{d}{dv}Q_2\right)\frac{Q_1}{Q_3} - \left(\frac{d}{dv}Q_3\right)\frac{Q_1Q_2}{Q_3^2}\right| \\
&\leq \left(\|\nabla_{\boldsymbol{\theta}}Q_1\| + \left|\frac{d}{dv}Q_1\right|\right)\left|\frac{Q_2}{Q_3}\right| + \left(\|\nabla_{\boldsymbol{\theta}}Q_2\| + \left|\frac{d}{dv}Q_2\right|\right)\left|\frac{Q_1}{Q_3}\right| + \left(\|\nabla_{\boldsymbol{\theta}}Q_3\| + \left|\frac{d}{dv}Q_3\right|\right)\left|\frac{Q_1Q_2}{Q_3^2}\right| \\
&\leq e^{k_{27}+k_{28}\|\boldsymbol{Y}\|},
\end{aligned}
$$

for some positive constants k_{27} and k_{28}. Therefore, by the mean-value theorem, we conclude that, for any $(v_1,\boldsymbol{\theta}_1,\Lambda_{s1})$ and $(v_2,\boldsymbol{\theta}_2,\Lambda_{s2})$ in $[0,\tau]\times\Theta\times\mathscr{A}$,

$$
\begin{aligned}
&\left|Q(v_1,\boldsymbol{O};\boldsymbol{\theta}_1,\Lambda_{s1}) - Q(v_2,\boldsymbol{O};\boldsymbol{\theta}_2,\Lambda_{s2})\right| \\
&\leq e^{k_{27}+k_{28}\|\boldsymbol{Y}\|}\left[\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\| + \left|\Lambda_{s1}(V) - \Lambda_{s2}(V)\right| + \int_0^\tau \left|\Lambda_{s1}(t) - \Lambda_{s2}(t)\right|dt + |v_1 - v_2|\right]
\end{aligned} \tag{4.34}
$$

holds for some positive constants k_{27} and k_{28} and $0 \leq V \leq \tau$ ($V = v_1$ or v_2).

Applying Theorem 2.7.5 (p159) in van der Vaart and Wellner (1996) to our situation, the entropy number for the class $\mathscr{A}$ satisfies $\log N_{[\cdot]}(\varepsilon, \mathscr{A}, L_2(P)) \leq K/\varepsilon$, where K is a constant. Thus, we can find $\exp\{K/\varepsilon\}$ brackets, $\{[L_j, U_j]\}$, to cover the class $\mathscr{A}$ such that $\|U_j - L_j\|_{L_2(P)} \leq \varepsilon$ for each pair of $[L_j, U_j]$. On the other hand, we can further find a partition of $[0,\tau] \times \Theta$, say $I_1 \bigcup I_2 \bigcup \cdots$, such that the number of partitions is of the order $(1/\varepsilon)^{d_\theta+1}$, and, for any $(v_1, \boldsymbol{\theta}_1)$ and $(v_2, \boldsymbol{\theta}_2)$ in the same partition, their Euclidean distance is less than ε. Therefore, the partition $\{I_1, I_2, \ldots\} \times \{[L_j, U_j]\}$ bracket covers $[0,\tau] \times \Theta \times \mathscr{A}$, and the total number of the partitions is of order $(1/\varepsilon)^{d_\theta+1} \exp\{1/\varepsilon\}$. Hence, from (4.34), for any I_l and $[L_j, U_j]$, the set of the functions $\{Q(v, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) : (v, \boldsymbol{\theta}) \in I_l, \Lambda_s \in \mathscr{A}, \Lambda_s \in [L_j, U_j]\}$ can be bracket covered by

$$\left[Q(v_l, \boldsymbol{O}; \boldsymbol{\theta}_l, \Lambda_{sl}) - e^{k_{27}+k_{28}\|\boldsymbol{Y}\|} \left\{ \varepsilon + \left| U_j(V) - L_j(V) \right| + \int_0^\tau \left| U_j(t) - L_j(t) \right| dt \right\}, \right.$$
$$\left. Q(v_l, \boldsymbol{O}; \boldsymbol{\theta}_l, \Lambda_{sl}) + e^{k_{27}+k_{28}\|\boldsymbol{Y}\|} \left\{ \varepsilon + \left| U_j(V) - L_j(V) \right| + \int_0^\tau \left| U_j(t) - L_j(t) \right| dt \right\} \right], \quad (4.35)$$

where $(v_l, \boldsymbol{\theta}_l)$ is a fixed point in I_l and Λ_{sj} is a fixed function in $[L_j, U_j]$. Note that the $L_2(P)$ distance between these two functions in the above bracket (4.35) is less than $O(\varepsilon)$. Therefore, we have

$$N_{[\cdot]}(\varepsilon, \mathscr{F}, \|\cdot\|_{L_2(P)} \leq O\left(\left(\frac{1}{\varepsilon} \right)^{d_\theta+1} e^{1/\varepsilon} \right).$$

Furthermore, $\mathscr{F}$ has an $L_2(P)$-integrable covering function, which is equal to $O($ $e^{k_{27}+k_{28}\|\boldsymbol{Y}\|})$. From Theorem 2.5.6 (p130) in van der Vaart and Wellner (1996), $\mathscr{F}$ is P-Donsker.

Additionally, in the above derivation, we also note that all the functions in $\mathscr{F}$ are bounded from below by $e^{-k_{29}-k_{30}\|\boldsymbol{Y}\|}$ for some positive constants k_{29} and k_{30}.

4.4.3.2 Proof of P-Donsker property of $\mathscr{G}$

Recall that we defined the class

$$\begin{aligned}\mathscr{G} \;=\; & \Big\{ l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0,\Lambda_{s0})^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}_0,\Lambda_{s0})[h_2], \\ & \qquad \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + \sup_{t\in[0,\tau]} |\Lambda_s(t)-\Lambda_{s0}(t)| \le \delta,\ (\mathbf{h}_1,h_2)\in\mathscr{H}\Big\},\end{aligned}$$

where $(\mathbf{h}_1^\beta, \mathbf{h}_1^\phi, \mathbf{h}_1^{\Sigma_b}, \mathbf{h}_1^\mu, \mathbf{h}_1^w, \mathbf{h}_1^\psi, \mathbf{h}_1^\gamma)$ denote the corresponding components of $\mathbf{h}_1$ for the parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathrm{Vec}(\boldsymbol{\Sigma}_b),\ \boldsymbol{\mu}, \boldsymbol{w}, \boldsymbol{\psi}, \boldsymbol{\gamma})$, respectively. We can write that for $(\mathbf{h}_1, h_2) \in \mathscr{H}$,

$$\begin{aligned}& l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2] \\ &= \left[\rho_1(\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 - \int_0^{V_s} \rho_2(t,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 d\Lambda_s(t)\right] + \Delta h_2(V_s) - \int_0^{V_s} \rho_3(t,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s) h_2(t) d\Lambda_s(t),\end{aligned}$$

where

$$\begin{aligned}& \rho_1(\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 \\ &= \left\{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s) d\boldsymbol{b}_\alpha\right\}^{-1} \times \sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s) \\ &\quad \times \Bigg[\sum_{j=1}^{n_N} \frac{1}{A(D(t_j;\phi))}\big(Y_j\boldsymbol{X}_j - B'(\boldsymbol{\beta};\boldsymbol{b}_\alpha)\big)\mathbf{h}_1^\beta \\ &\qquad + \sum_{j=1}^{n_N}\left\{-\left(\frac{Y_j(\boldsymbol{X}_j\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_j\boldsymbol{b}_\alpha) - B(\boldsymbol{\beta};\boldsymbol{b}_\alpha)}{A(D(t_j;\phi))^2}\right)A'(D(t_j;\phi)) + C'(Y_j;D(t_j;\phi))\right\}\mathrm{h}_1^\phi \\ &\qquad + \frac{1}{2}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_\alpha)^T\boldsymbol{\Sigma}_b^{-1}\boldsymbol{D}_b\boldsymbol{\Sigma}_b^{-1}(\boldsymbol{b}_\alpha-\boldsymbol{\mu}_\alpha) - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_b^{-1}\boldsymbol{D}_b\right) \\ &\qquad + \left(\boldsymbol{b}_\alpha - \frac{1}{2}\boldsymbol{\mu}_\alpha\right)^T\boldsymbol{\Sigma}_b^{-1}\mathbf{h}_1^{\mu_\alpha} + \frac{1}{w_\alpha}\mathbf{h}_1^{w_\alpha} + \Delta_s\big\{\big(\widetilde{\boldsymbol{Z}}(V_s)\circ\boldsymbol{b}_\alpha^T\big)\mathbf{h}_1^\psi + \boldsymbol{Z}(V_s)\mathbf{h}_1^\gamma\big\}\Bigg] d\boldsymbol{b}_\alpha, \\ & \rho_2(t,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1 \\ &= \left\{\sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s) d\boldsymbol{b}_\alpha\right\}^{-1} \times \sum_\alpha \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b},\alpha,\boldsymbol{O};\boldsymbol{\theta},\Lambda_s) \\ &\quad \times \exp\left\{\widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi}\circ\boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma}\right\} \times \left[\big(\widetilde{\boldsymbol{Z}}(t)\circ\boldsymbol{b}_\alpha^T\big)\mathbf{h}_1^\psi + \boldsymbol{Z}(t)\mathbf{h}_1^\gamma\right] d\boldsymbol{b}_\alpha,\end{aligned}$$

$$
\begin{aligned}
&\rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) \\
&= \left\{ \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) d\boldsymbol{b}_\alpha \right\}^{-1} \\
&\quad \times \sum_{\alpha} \int_{\boldsymbol{b}_\alpha} G(\boldsymbol{b}, \alpha, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) \times \exp\left\{ \widetilde{\boldsymbol{Z}}(t)(\boldsymbol{\psi} \circ \boldsymbol{b}_\alpha) + \boldsymbol{Z}(t)\boldsymbol{\gamma} \right\} d\boldsymbol{b}_\alpha,
\end{aligned}
$$

$B'(\boldsymbol{\beta}; \boldsymbol{b})$ is the derivative of $B(\boldsymbol{\beta}; \boldsymbol{b})$ with respect to $\boldsymbol{\beta}$, $A'(D(t_j; \phi))$ and $C'(Y_j; D(t_j; \phi))$ are the derivatives of $A(D(t_j; \phi))$ and $C(Y_j; D(t_j; \phi))$ with respect to ϕ respectively, and $\boldsymbol{D}_b$ is the symmetric matrix such that $\mathrm{Vec}(\boldsymbol{D}_b) = \mathbf{h}_1^b$.

For $l = 1, 2, 3$, we denote $\nabla_{\boldsymbol{\theta}} \rho_l$ and $\nabla_{\Lambda_s} \rho_l[\delta\Lambda_s]$ as the derivatives of ρ_l with respect to $\boldsymbol{\theta}$ and Λ_s along the path $\Lambda_s + \varepsilon\delta\Lambda_s$. Then, using the similar arguments done in Section 4.4.3.1, it is verified that $\nabla_{\Lambda_s} \rho_l[\delta\Lambda_s] = \int_0^t \rho_{l+3}(u, \boldsymbol{O}; \boldsymbol{\theta}, \Lambda_s) d\delta\Lambda_s(u)$ and there exist two positive constants q_1 and q_2 such that

$$
\sum_{l} \left\{ |\rho_l| + |\nabla_{\boldsymbol{\theta}} \rho_l| \right\} \leq e^{q_1 + q_2 \|\boldsymbol{Y}\|}
$$

By the mean value theorem, we have that, for any $(\boldsymbol{\theta}, \Lambda_s, \mathbf{h}_1, h_2)$ and $(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s, \widetilde{\mathbf{h}}_1, \widetilde{h}_2)$ in $\Xi \times \mathscr{H}$,

$$
\begin{aligned}
&l_\theta(\boldsymbol{\theta}, \Lambda_s)^T \mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)^T \widetilde{\mathbf{h}}_1 - l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)[\widetilde{h}_2] \\
&= l_\theta(\boldsymbol{\theta}, \Lambda_s)^T \mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)^T \mathbf{h}_1 - l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)[h_2] \\
&\quad + l_\theta(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)^T \mathbf{h}_1 + l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)[h_2] - l_\theta(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)^T \widetilde{\mathbf{h}}_1 - l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)[\widetilde{h}_2] \\
&= \left[l_\theta(\boldsymbol{\theta}, \Lambda_s)^T - l_\theta(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)^T \right] \mathbf{h}_1 + \left[l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s) - l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s) \right][h_2] \\
&\quad + l_\theta(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)^T (\mathbf{h}_1 - \widetilde{\mathbf{h}}_1) + l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)([h_2] - [\widetilde{h}_2]) \\
&= (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^T \left[\frac{d}{d\boldsymbol{\theta}} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \Lambda_s) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*, \Lambda_s = \Lambda_s^*} \right] \mathbf{h}_1 + \left[\frac{d}{d\Lambda_s} l_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \Lambda_s) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*, \Lambda_s = \Lambda_s^*} \right]^T [\Lambda_s - \widetilde{\Lambda}_s] \mathbf{h}_1 \\
&\quad + (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^T \left[\frac{d}{d\boldsymbol{\theta}} l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*, \Lambda_s = \Lambda_s^*} \right] [h_2] + \left[\frac{d}{d\Lambda_s} l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s) \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}^*, \Lambda_s = \Lambda_s^*} \right]^T [\Lambda_s - \widetilde{\Lambda}_s][h_2] \\
&\quad + l_\theta(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)^T (\mathbf{h}_1 - \widetilde{\mathbf{h}}_1) + l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}}, \widetilde{\Lambda}_s)([h_2] - [\widetilde{h}_2])
\end{aligned}
$$

$$
\begin{aligned}
&= (\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^T\nabla_{\boldsymbol{\theta}}\rho_1(\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)\mathbf{h}_1 - (\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^T\int_0^{V_S}\nabla_{\boldsymbol{\theta}}\rho_2(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T d\Lambda_s^*(t)\mathbf{h}_1\\
&\quad+\int_0^{V_s}\rho_4(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T\mathbf{h}_1 d(\Lambda_s-\widetilde{\Lambda}_s)(t)\\
&\quad+\int_0^{V_s}\int_0^t\rho_5(u,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T d(\Lambda_s-\widetilde{\Lambda}_s)(u)\mathbf{h}_1 d\Lambda_s^*(t)\\
&\quad-\int_0^{V_s}\rho_2(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T(\Lambda_s-\widetilde{\Lambda}_s)\mathbf{h}_1 dt - (\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^T\int_0^{V_S}\nabla_{\boldsymbol{\theta}}\rho_3(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)h_2(t)d\Lambda_s^*(t)\\
&\quad+\int_0^{V_s}\int_0^t\rho_6(u,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)d(\Lambda_s-\widetilde{\Lambda}_s)(u)h_2(t)d\Lambda_s^*(t)\\
&\quad-\int_0^{V_s}\rho_3(t,\boldsymbol{O};\boldsymbol{\theta}^*,\Lambda_s^*)^T(\Lambda_s-\widetilde{\Lambda}_s)(t)h_2(t)dt\\
&\quad+\rho_1(\boldsymbol{O};\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T(\mathbf{h}_1-\widetilde{\mathbf{h}}_1)-\int_0^{V_S}\rho_2(t,\boldsymbol{O};\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T(\mathbf{h}_1-\widetilde{\mathbf{h}}_1)d\widetilde{\Lambda}_s(t)\\
&\quad+\Delta_s(h_2(V_s)-\widetilde{h}_2(V_s))-\int_0^{V_S}\rho_3(t,\boldsymbol{O};\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)(h_2(V_s)-\widetilde{h}_2(V_s))d\widetilde{\Lambda}_s(t),
\end{aligned}
\tag{4.36}
$$

where $(\boldsymbol{\theta}^*,\Lambda_s^*)$ is equal to $\varepsilon^*(\boldsymbol{\theta},\Lambda_s)+(1-\varepsilon^*)(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)$ for some $\varepsilon^*\in[0,1]$. Thus, we have

$$
\begin{aligned}
&\left|l_\theta(\boldsymbol{\theta},\Lambda_s)^T\mathbf{h}_1+l_{\Lambda_s}(\boldsymbol{\theta},\Lambda_s)[h_2]-l_\theta(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)^T\widetilde{\mathbf{h}}_1-l_{\Lambda_s}(\widetilde{\boldsymbol{\theta}},\widetilde{\Lambda}_s)[\widetilde{h}_2]\right|\\
&\le e^{q_1+q_2\|\boldsymbol{Y}\|}\Bigg\{\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\|+\|\mathbf{h}_1-\widetilde{\mathbf{h}}_1\|+|\Lambda_s(V_s)-\widetilde{\Lambda}_s(V_s)|\\
&\qquad+\int_0^\tau|\Lambda_s(t)-\widetilde{\Lambda}_s(t)|[dt+d|h_2(t)|+d|\widetilde{h}_2(t)|]\\
&\qquad+|h_2(V_s)-\widetilde{h}_2(V_s)|+\int_0^\tau|h_2(V_s)-\widetilde{h}_2(V_s)|[d\Lambda_s(t)-d\widetilde{\Lambda}_s(t)]\Bigg\},
\end{aligned}
\tag{4.37}
$$

where $d|h_2(t)| = dh_2^+(t)+dh_2^-(t)$ and $d|\widetilde{h}_2(t)| = d\widetilde{h}_2^+(t)+d\widetilde{h}_2^-(t)$. As done in Section 4.4.3.1, by applying Theorem 2.7.5 (p159) in van der Vaart and Wellner (1996), we note that for a set $\mathscr{H}_2=\{h_2:\|h_2\|_V\le B_1\}$, $\log N_{[\cdot]}\big(\varepsilon,\mathscr{H}_2,L_2(P)\big)\le K/\varepsilon$ for a constant B_1 and any probability measure P where K is a constant. Thus, we can find $\exp\{K/\varepsilon\}$ brackets, $\{[L_j,U_j]\}$, to cover the class $\mathscr{H}_2$ such that $\|U_j-L_j\|_{L_2(P)}\le\varepsilon$ for each pair of $[L_j,U_j]$. On the other hand, we can further find a partition of $\mathscr{H}_1=\{\mathbf{h}_1:\|\mathbf{h}_1\|\le 1\}$, say $I_1\bigcup I_2\bigcup\cdots$, such that the number of partitions is of the order $(1/\varepsilon)$, and, for any $\mathbf{h}_1$ and h_2 in the same partition, their Euclidean distance is less than ε. Therefore, the

partition $\{I_1, I_2, \ldots\} \times \{[L_j, U_j]\}$ bracket covers $\mathscr{H}_1 \times \mathscr{H}_2$, and the total number of the partitions is of order $(1/\varepsilon)\exp\{1/\varepsilon\}$. Then, we obtain

$$\log N_{[\cdot]}\big(\varepsilon, \mathscr{G}, L_2(P)\big) \leq O\bigg(\frac{1}{\varepsilon} + \log\varepsilon\bigg).$$

Moreover, $\mathscr{G}$ has an $L_2(P)$-integrable covering function, which is equal to $O(e^{q_1+q_2\|\boldsymbol{Y}\|})$. Hence, from Theorem 2.5.6 (p130) in van der Vaart and Wellner (1996), $\mathscr{G}$ is P-Donsker.

Additionally, from (4.37), we can calculate

$$\begin{aligned}
&\Big|l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2]\Big| \\
&\leq e^{q_1+q_2\|\boldsymbol{Y}\|}\bigg\{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + |\Lambda_s(V_s) - \Lambda_{s0}(V_s)| + \int_0^\tau |\Lambda_s(t) - \Lambda_{s0}(t)|dt\bigg\} \\
&\quad + \bigg|\int_0^\tau \rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)h_2(t)d(\Lambda_s(t) - \Lambda_{s0}(t))\bigg|. \qquad (4.38)
\end{aligned}$$

If $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \to 0$ and $\sup_{t\in[0,\tau]}|\Lambda_s(t) - \Lambda_{s0}(t)| \to 0$, the above expression converges to zero uniformly. Therefore,

$$\sup_{(\mathbf{h}_1, h_2)\in\mathscr{H}} \mathbf{P}\big[l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2]\big]^2 \longrightarrow 0.$$

4.4.3.3 Derivative operator ∇U_{ξ_0}

From (4.36) in the previous Section 4.4.3.2, we can obtain

$$\begin{aligned}
&l_\theta(\boldsymbol{\theta}, \Lambda_s)^T\mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\mathbf{h}_1 - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2] \\
&= \big[l_\theta(\boldsymbol{\theta}, \Lambda_s)^T - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T\big]\mathbf{h}_1 + \big[l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s) - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})\big][h_2] \\
&= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}}\rho_1(\boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)\mathbf{h}_1 - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\int_0^{V_S} \nabla_{\boldsymbol{\theta}}\rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)^T\mathbf{h}_1 d\Lambda_s^*(t) \\
&\quad + \int_0^{V_s} \rho_4(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)^T\mathbf{h}_1 d(\Lambda_s - \Lambda_{s0})(t) \\
&\quad + \int_0^{V_s}\int_0^t \rho_5(u, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)^T d(\Lambda_s - \Lambda_{s0})(u)\mathbf{h}_1 d\Lambda_s^*(t)
\end{aligned}$$

$$
\begin{aligned}
&- \int_0^{V_s} \rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)^T \mathbf{h}_1 d(\Lambda_s - \Lambda_{s0})(t) \\
&- (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) h_2(t) d\Lambda_s^*(t) \\
&+ \int_0^{V_s} \int_0^t \rho_6(u, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)^T d(\Lambda_s - \Lambda_{s0})(u) h_2(t) d\Lambda_s^*(t) \\
&- \int_0^{V_s} \rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)^T h_2(t) d(\Lambda_s - \Lambda_{s0})(t) \\
= &(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left\{ \nabla_{\boldsymbol{\theta}} \rho_1(\boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) - \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*)^T d\Lambda_s^*(t) \right\} \mathbf{h}_1 \\
&+ \mathbf{h}_1^T \left\{ \int_0^\tau I(t \le V_s) \big[\rho_4(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) - \rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) \right. \\
&\qquad \left. + \rho_5(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) \int_t^{V_s} d\Lambda_s^*(u) \big] d(\Lambda_s - \Lambda_{s0})(t) \right\} \\
&- (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \int_0^\tau I(t \le V_s) \nabla_{\boldsymbol{\theta}} \rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) h_2(t) d\Lambda_s^*(t) \\
&- \int_0^\tau \left\{ - I(t \le V_s) \rho_6(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) \int_t^{V_s} h_2(u) d\Lambda_s^*(u) \right. \\
&\qquad \left. + I(t \le V_s) \rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) h_2(t) \right\} d(\Lambda_s - \Lambda_{s0})(t). \qquad (4.39)
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\nabla U_{\xi_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0, \Lambda_s - \Lambda_{s0})[\mathbf{h}_1, h_2] \\
= &(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{P} \left\{ \nabla_{\boldsymbol{\theta}} \rho_1(\boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\Lambda_{s0}(t) \right\} \mathbf{h}_1 \\
&+ \mathbf{h}_1^T \left\{ \int_0^\tau \mathbf{P} \left[I(t \le V_s) \left(\rho_4(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \right. \right. \right. \\
&\qquad \left. \left. \left. + \rho_5(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_t^{V_s} d\Lambda_{s0}(u) \right) \right] d(\Lambda_s - \Lambda_{s0})(t) \right\} \\
&- (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \int_0^\tau \mathbf{P} \left\{ I(t \le V_s) \nabla_{\boldsymbol{\theta}} \rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \right\} h_2(t) d\Lambda_{s0}(t) \\
&- \int_0^\tau \mathbf{P} \left\{ - I(t \le V_s) \rho_6(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_t^{V_s} h_2(u) d\Lambda_{s0}(u) \right. \\
&\qquad \left. + I(t \le V_s) \rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) h_2(t) \right\} d(\Lambda_s - \Lambda_{s0})(t).
\end{aligned}
$$

By the similar algebra done in (4.38), we can verify that, for $j = 1, \ldots, 6$,

$$\sup_{t\epsilon[0,\tau]} \|\rho_j(t, \boldsymbol{O}; \boldsymbol{\theta}^*, \Lambda_s^*) - \rho_j(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\| \leq e^{q_3+q_4\|\boldsymbol{Y}\|}\left\{\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\| + \sup_{t\epsilon[0,\tau]} |\Lambda_s^* - \Lambda_{s0}|\right\},$$

which implies that the linear operator ∇U_{ξ_0} is bounded.

Then, we obtain

$$\begin{aligned}
&\mathbf{P}\left[l_\theta(\boldsymbol{\theta}, \Lambda_s)^T \mathbf{h}_1 + l_{\Lambda_s}(\boldsymbol{\theta}, \Lambda_s)[h_2] - l_\theta(\boldsymbol{\theta}_0, \Lambda_{s0})^T \mathbf{h}_1 - l_{\Lambda_s}(\boldsymbol{\theta}_0, \Lambda_{s0})[h_2]\right] \\
&= \nabla U_{\xi_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0, \Lambda_s - \Lambda_{s0})[\mathbf{h}_1, h_2] + o\Big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \sup_{t\epsilon[0,\tau]} |\Lambda_s - \Lambda_{s0}|\Big)(\|\mathbf{h}_1\| + \|h_2\|_V).
\end{aligned}$$

Therefore, U_ξ is *Fréchet* differentiable at ξ_0.

Additionally, from (4.39) and the above expression, we have

$$\nabla U_{\xi_0}(\boldsymbol{\theta} - \boldsymbol{\theta}_0, \Lambda_s - \Lambda_{s0})[\mathbf{h}_1, h_2] = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \Omega_1[\mathbf{h}_1, h_2] + \int_0^\tau \Omega_2[\mathbf{h}_1, h_2] d(\Lambda_s - \Lambda_{s0})(t),$$

where

$$\begin{aligned}
\Omega_1[\mathbf{h}_1, h_2] \;=\; & \mathbf{P}\left\{\nabla_{\boldsymbol{\theta}} \rho_1(\boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \int_0^{V_S} \nabla_{\boldsymbol{\theta}} \rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) d\Lambda_{s0}(t)\right\}\mathbf{h}_1 \\
& - \int_0^\tau \mathbf{P}\left\{I(t \leq V_s) \nabla_{\boldsymbol{\theta}} \rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\} h_2(t) d\Lambda_{s0}(t)
\end{aligned}$$

and

$$\begin{aligned}
&\Omega_2[\mathbf{h}_1, h_2] \\
&= \mathbf{h}_1^T \mathbf{P}\left\{I(t \leq V_s)\big[\rho_4(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) + \rho_5(u, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_t^{V_s} d\Lambda_{s0}(u)\big]\right\} \\
&\quad + \mathbf{P}\left\{I(t \leq V_s)\rho_6(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_t^{V_s} h_2(u) d\Lambda_{s0}(u)\right\} \\
&\quad - \mathbf{P}\left\{I(t \leq V_s)\rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\} h_2(t).
\end{aligned}$$

Thus, $\Omega = (\Omega_1, \Omega_2)$ is the bounded linear operator from $R^d \times BV[0,\tau]$ to itself. Furthermore, we note that $\Omega = \boldsymbol{H} + (\boldsymbol{M}_1, \boldsymbol{M}_2)$, where

$$
\begin{aligned}
\boldsymbol{H}(\mathbf{h}_1, h_2) &= \Big(\mathbf{h}_1, -\mathbf{P}\left\{I(t \le V_s)\rho_3(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0})\right\}h_2(t)\Big), \\
\boldsymbol{M}_1(\mathbf{h}_1, h_2) &= \Omega_1[\mathbf{h}_1, h_2] - \mathbf{h}_1, \\
\boldsymbol{M}_2(\mathbf{h}_1, h_2) &= \mathbf{h}_1^T \mathbf{P}\left\{I(t \le V_s)\big[\nabla_{\boldsymbol{\theta}}\rho_4(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) - \rho_2(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \right.\right. \\
&\qquad \left.\left. + \rho_5(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_t^{V_s} d\Lambda_{s0}(u)\big]\right\} \\
&\qquad + \mathbf{P}\left\{I(t \le V_s)\rho_6(t, \boldsymbol{O}; \boldsymbol{\theta}_0, \Lambda_{s0}) \int_t^{V_s} h_2(u) d\Lambda_{s0}(u)\right\},
\end{aligned}
$$

and also note that $\boldsymbol{H}$ is obviously invertible. Since $\boldsymbol{M}_1$ maps into a finite-dimensional space, it is compact. The image of $\boldsymbol{M}_2$ is a continuously differentiable function in $[0,\tau]$. By the Arzela-Ascoli theorem (p41) in van der Vaart and Wellner (1996), $\boldsymbol{M}_2$ is a compact operator from $R^d \times BV[0,\tau]$ to $BV[0,\tau]$. Thus, we conclude that Ω is the summation of an invertible operator $\boldsymbol{H}$ and a compact operator $\boldsymbol{M} = (\boldsymbol{M}_1, \boldsymbol{M}_2)$.

4.5 Simulation Studies

In this section, we present the results from our simulation studies. First, to assess finite sample properties of the proposed maximum likelihood estimators, two sets of simulations with different generalized linear mixed models for the longitudinal outcomes are performed. Continuous and binary data are considered for longitudinal process in the simulations in Sections 4.5.1 and 4.5.2, respectively. Then, we conduct simulation studies for robustness of the assumed mixture distribution in Section 4.5.3. Selection procedures for the number of mixtures by AIC and BIC criteria are assessed through simulation studies in Section 4.5.4.

4.5.1 Continuous longitudinal outcomes and survival time

In this section, we assume Y_{ij} follows a Gaussian distribution given a subject-specific random intercept. Specifically we have

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + b_i + \epsilon_{ij} = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3ij} + b_i + \epsilon_{ij},$$

for $j = 1, \ldots, n_i$, where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2)$, and

$$h(t|b_i) = \lambda(t)\exp\{\psi b_i + \boldsymbol{Z}_i(t)\boldsymbol{\gamma}\} = \lambda(t)\exp\{\psi b_i + \gamma_1 Z_{1i} + \gamma_2 Z_{2i}\},$$

where $b_i \sim \sum_{k=1}^{K} w_k \mathcal{N}(\mu_k, \sigma_b^2)$, K is the number of mixture components, and $K = 2$ and $K = 3$ are simulated. $X_{1i} \equiv Z_{1i}$ are simulated from a Bernoulli distribution with success probability being 0.5, and $X_{2i} \equiv Z_{2i}$ are simulated from the uniform distribution between 0 and 1. The longitudinal data are generated for every 0.1 unit of time, and thus X_{3ij}, the time at measurement, has the value of every 0.1 unit ranging over 0 through 2.4. We consider $\psi = -0.1$ indicating negative dependency between longitudinal process and survival time model. The parameters in the longitudinal and hazard models are chosen as $\beta_1 = 1$, $\beta_2 = -0.5$, $\beta_3 = -0.2$, $\sigma_y^2 = 0.5$, $\psi = -0.1$, $\gamma_1 = -0.1$, $\gamma_2 = 0.1$, and $\lambda(t) = 1$. The parameters in the mixture distribution for random effects are $\mu_1 = -1.5$, $\mu_2 = 1.5$, and $w_1 = 0.4$ for $K = 2$ and $\mu_1 = -3$, $\mu_2 = 0$, $\mu_3 = 3$, $w_1 = 0.4$, and $w_2 = 0.3$ for $K = 3$. The weight of the last mixture component (w_2 and w_3 for $K = 2$ and $K = 3$ respectively) is determined from the restriction $\sum_{k=1}^{K} w_k = 1$. The variance of random effects σ_b^2 is chosen as 0.3. Censoring time is generated from the uniform distribution between 0.4 and 2.4, and the censoring proportion is around 25~35%. We consider different sample sizes (n=400, 800) with 1000 replications. The average number of longitudinal observations (n_i) is 7–8 with the range of 1 to 24. For the estimated baseline cumulative hazard function, we consider three fixed time points of 0.9, 1.4, and 1.9.

The results of the maximum likelihood estimates for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma_y^2, \boldsymbol{\mu}^T, \boldsymbol{w}^T, \sigma_b^2, \psi, \boldsymbol{\gamma}^T)^T$ and baseline cumulative hazards at the three time points and their respective standard error estimates are reported in Table 4.1. In Table 4.1, "True" gives the true values of parameters; the averages of the maximum likelihood estimates from the EM algorithm are in "Est."; the sample standard deviations from 1000 simulations are reported in "SSD"; "ESE" is the average of 1000 standard error estimates based on the observed information matrix; "CP" is the coverage proportion of 95% confidence intervals based on the estimated standard error "ESE". Satterthwaite method is used for the coverage probabilities of σ_y^2 and σ_b^2.

From Table 4.1, we can see that even for the smaller sample size (n=400), the bias of the estimates from EM algorithm is negligible for most cases. The estimated standard errors calculated from the observed information matrix are close to the sample standard deviations from the 1000 estimates, and the 95% confidence interval coverage rates are close to 0.95 except for weights of the mixture components. The coverage rates of weights are improved for larger sample size in both 2 and 3 mixtures. The estimates for the parameters in the longitudinal and hazards models ($\boldsymbol{\beta}$, σ_y^2, ψ, $\boldsymbol{\gamma}$ and $\Lambda(t)$) perform well for different mixtures.

4.5.2 Binary longitudinal outcomes and survival time

In this section, we assume that Y_{ij} is a binary outcome following

$$P(Y_{ij} = y_{ij}|b_i) = \exp\left\{y_{ij}\eta_{ij} - \log(1 + \exp\{\eta_{ij}\})\right\}, \qquad y_{ij} = 0,\ 1,$$

with $\eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + b_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3ij} + b_i$ for $j = 1, \ldots, n_i$, and we consider the same hazards model and simulation setting as those used in Section 4.5.1 except the followings. The parameters in the mixture distribution for random effects are $\mu_1 = -3$,

Table 4.1: Summary of simulation results of maximum likelihood estimation using mixtures of Gaussian distributions for random effects in the joint modeling of continuous longitudinal outcomes and survival time.

			n=400				n=800			
Mixture	Par.	True	Est.	SSD	ESE	CP	Est.	SSD	ESE	CP
2	β_1	1.0	.983	.066	.068	.958	.985	.047	.048	.947
	β_2	- .5	- .529	.107	.119	.969	- .540	.079	.084	.947
	β_3	- .2	- .203	.033	.033	.955	- .203	.024	.024	.952
	σ_y^2	.5	.500	.014	.014	.954	.500	.010	.010	.948
	μ_1	-1.5	-1.478	.081	.088	.962	-1.469	.060	.062	.938
	μ_2	1.5	1.524	.075	.082	.966	1.530	.055	.058	.940
	w_1	.4	.400	.025	.033	.991	.401	.018	.023	.981
	σ_b^2	.3	.296	.029	.029	.955	.298	.020	.020	.958
	ψ	- .1	- .102	.040	.039	.950	- .100	.028	.028	.946
	γ_1	- .1	- .101	.123	.121	.945	- .105	.085	.085	.952
	γ_2	.1	.102	.209	.210	.954	.096	.144	.147	.950
	$\Lambda(\ .9)$	.9	.911	.130	.128	.950	.909	.087	.090	.955
	$\Lambda(1.4)$	1.4	1.421	.206	.202	.942	1.415	.139	.141	.952
	$\Lambda(1.9)$	1.9	1.939	.304	.295	.953	1.924	.205	.205	.950
3	β_1	1.0	.983	.070	.071	.947	.984	.049	.050	.956
	β_2	- .5	- .543	.116	.123	.952	- .543	.085	.087	.922
	β_3	- .2	- .203	.034	.034	.949	- .204	.024	.024	.960
	σ_y^2	.5	.500	.014	.014	.957	.500	.010	.010	.950
	μ_1	-3.0	-2.970	.084	.090	.954	-2.968	.064	.063	.909
	μ_2	.0	.028	.093	.097	.954	.032	.069	.068	.933
	μ_3	3.0	3.030	.089	.094	.954	3.034	.063	.066	.925
	w_1	.4	.400	.025	.033	.992	.400	.018	.023	.983
	w_2	.3	.299	.024	.029	.980	.300	.017	.020	.977
	σ_b^2	.3	.295	.029	.029	.956	.298	.021	.021	.946
	ψ	- .1	- .101	.024	.024	.956	- .101	.017	.017	.941
	γ_1	- .1	- .091	.112	.119	.963	- .096	.085	.084	.950
	γ_2	.1	.088	.215	.207	.946	.114	.146	.146	.944
	$\Lambda(\ .9)$	.9	.913	.125	.127	.948	.897	.088	.088	.951
	$\Lambda(1.4)$	1.4	1.417	.202	.200	.949	1.402	.141	.140	.949
	$\Lambda(1.9)$	1.9	1.928	.297	.292	.946	1.908	.206	.204	.948

$\mu_2 = 3$, and $w_1 = 0.4$ for $K = 2$ and $\mu_1 = -6$, $\mu_2 = 0$, $\mu_3 = 6$, $w_1 = 0.4$, and $w_2 = 0.3$ for $K = 3$. The binary longitudinal data are generated for every 0.1 and 0.05 units of time for the mixture of 2 and 3 normal distributions, respectively, and X_{3ij}, the time at measurement, has the values of every 0.1 and 0.05 units corresponding to the mixture distributions ranging over 0 through 2.4. Thus, the average numbers of longitudinal observations (n_i) are 7–8 with the range of 1 to 24 and 15–16 with the range of 1 to 48 for the mixture of 2 and 3 distributions, respectively.

The results of the maximum likelihood estimates for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\mu}^T, \boldsymbol{w}^T, \sigma_b^2, \psi, \boldsymbol{\gamma}^T)^T$ and baseline cumulative hazards at the given three time points and their respective standard error estimates are reported in Table 4.2. Similar to the results for the continuous longitudinal outcomes, Table 4.2 shows that overall the estimates perform well even for the smaller sample size $n = 400$ with small biases. The parameters of interest in longitudinal and hazards models have the estimated standard errors which are close to the sample standard deviations. Meanwhile, the estimated standard errors of the parameters of mixture components which are means of random effects and weights appear to be overestimated being larger than their sample standard deviations, which leads to the wide confidence interval.

4.5.3 Sensitivity for model-misspecification

In this section, we conduct simulation studies to examine the sensitivity of the assumed mixture distribution. We consider continuous longitudinal outcomes and survival time with the same setting used in Section 4.5.1 except for the true distribution of random effects. Random effects are generated from a mixture of a t-distribution with 10 degrees of freedom and non-centrality of -1 and a Gamma distribution with shape and scale parameters of 7 and 1/8 respectively. We assume equal probability for the two distributions. We fit 5 sets of simultaneous models assuming different mixtures for random

Table 4.2: Summary of simulation results of maximum likelihood estimation using mixtures of Gaussian distributions for random effects in the joint modeling of binary longitudinal outcomes and survival time.

			n=400				n=800			
Mixture	Par.	True	Est.	SSD	ESE	CP	Est.	SSD	ESE	CP
2	β_1	1.0	1.029	.193	.201	.960	1.015	.143	.141	.942
	β_2	- .5	- .508	.292	.323	.966	- .495	.205	.227	.965
	β_3	- .2	- .200	.166	.180	.966	- .203	.116	.127	.968
	μ_1	-3.0	-3.046	.241	.275	.968	-3.034	.164	.193	.970
	μ_2	3.0	3.016	.211	.253	.976	3.011	.142	.177	.984
	w_1	.4	.401	.025	.033	.993	.400	.017	.023	.991
	σ_b^2	.3	.329	.133	.195	.940	.332	.092	.136	.956
	ψ	- .1	- .099	.021	.021	.949	- .099	.015	.015	.955
	γ_1	- .1	- .103	.121	.122	.959	- .098	.087	.086	.947
	γ_2	.1	.091	.210	.211	.944	.104	.142	.149	.958
	$\Lambda(\ .9)$	.9	.910	.131	.130	.955	.900	.088	.091	.956
	$\Lambda(1.4)$	1.4	1.421	.209	.206	.934	1.402	.142	.143	.956
	$\Lambda(1.9)$	1.9	1.932	.310	.299	.941	1.899	.205	.207	.948
3	β_1	1.0	.988	.167	.171	.953	.993	.123	.121	.947
	β_2	- .5	- .519	.268	.287	.960	- .516	.189	.203	.967
	β_3	- .2	- .208	.126	.128	.957	- .206	.091	.091	.951
	μ_1	-6.0	-5.844	.353	.483	.967	-5.872	.260	.342	.963
	μ_2	.0	.023	.172	.194	.970	.018	.127	.138	.966
	μ_3	6.0	6.024	.397	.504	.984	6.006	.303	.349	.971
	w_1	.4	.402	.025	.035	.995	.402	.018	.024	.989
	w_2	.3	.298	.025	.034	.986	.298	.017	.024	.985
	σ_b^2	.3	.277	.095	.100	.977	.289	.070	.072	.966
	ψ	- .1	- .102	.014	.015	.955	- .101	.011	.010	.946
	γ_1	- .1	- .103	.121	.120	.955	- .107	.085	.084	.948
	γ_2	.1	.104	.201	.208	.961	.099	.147	.146	.949
	$\Lambda(\ .9)$	.9	.909	.128	.130	.950	.911	.094	.092	.930
	$\Lambda(1.4)$	1.4	1.421	.202	.207	.960	1.420	.147	.146	.946
	$\Lambda(1.9)$	1.9	1.926	.297	.302	.958	1.929	.220	.213	.946

effects which are 1 normal distribution without mixture and the mixtures of 2, 3, 4 and 5 normal distributions, and we compare the results for the parameters of interest in longitudinal and hazards models and the estimated density plots of random effects. Table 4.3 shows the results of longitudinal and hazards models from assuming the 5 different models for random effects. We can see that bias gets smaller and coverage rates become closer to the 95% nominal level as the number of mixtures increases. From the table, we also find that more mixture produces estimates more close to the true values in the longitudinal model while estimates in hazards model are less sensitive to the number of distributions in mixture. In other words, when the true distribution of random effects is not a Gaussian distribution, the use of mixture is effective in longitudinal model but the inference on hazards model is reasonable regardless of mixture. Figure 4.1 shows the true and estimated density plots of random effects. From these density plots, all the mixture models of 2, 3, 4 and 5 normal distributions produces similar shapes to the true distribution while one normal distribution does not. The mixture of 5 normal distribution appears to be close to the true density. Figure 4.2 shows the relative bias plot of the parameters in longitudinal and hazard models which are denoted with thin and thick lines respectively. The relative biases are calculated from the median absolute biases divided by their absolute true values. This Figure 4.2 confirms what we observe in Table 4.3.

4.5.4 Selection of the number of mixture distributions

We adopt Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) for selection of the number of normal distribution in mixture and assess these selection procedures through simulation studies in this section. AIC gives a penalty to a model with more parameters and BIC gives a penalty to a model with more parameters and larger sample size. Given a data set, competing models are ranked according to their

Table 4.3: Summary of simulation results of sensitivity for model-misspecification

		1 Normal distribution				2 Normal distributions				3 Normal distributions				4 Normal distributions				5 Normal distributions			
Par.	TRUE	Est.	SSD	ESE	CP	Est.	SSD	ESE	CP	Est.	SSD	ESE	CP	Est.	SSD	ESE	CP	Est.	SSD	ESE	CP
< Longitudinal model >																					
β_1	1.0	.947	.085	.095	.931	.965	.074	.074	.928	.977	.061	.060	.935	.978	.061	.060	.934	.985	.055	.054	.943
β_2	- .5	- .591	.122	.164	.975	- .582	.118	.128	.933	- .557	.101	.105	.930	- .555	.101	.104	.928	- .536	.092	.095	.928
β_3	- .2	- .203	.024	.025	.964	- .202	.024	.025	.961	- .202	.024	.024	.961	- .202	.024	.024	.959	- .202	.024	.024	.955
σ_y^2	.5	.501	.010	.010	.948	.501	.010	.010	.948	.501	.010	.010	.950	.501	.010	.010	.952	.500	.010	.010	.946
< Hazards model >																					
ψ	- .1	- .102	.034	.033	.942	- .101	.034	.033	.943	- .101	.034	.033	.941	- .101	.034	.033	.940	- .102	.034	.033	.945
γ_1	- .1	- .093	.083	.085	.949	- .095	.083	.084	.953	- .096	.083	.084	.954	- .096	.083	.084	.951	- .098	.085	.084	.948
γ_2	.1	.107	.144	.147	.949	.105	.144	.147	.949	.103	.143	.146	.951	.102	.143	.146	.951	.108	.144	.146	.952
$\Lambda(.9)$	.9	.906	.089	.089	.943	.907	.089	.089	.944	.906	.089	.089	.945	.907	.089	.089	.943	.904	.090	.089	.944
$\Lambda(1.4)$	1.4	1.408	.139	.140	.945	1.409	.139	.140	.945	1.408	.139	.140	.947	1.409	.139	.140	.946	1.403	.137	.139	.947
$\Lambda(1.9)$	1.9	1.911	.202	.203	.956	1.911	.202	.203	.956	1.910	.200	.203	.958	1.911	.202	.203	.956	1.905	.200	.202	.954

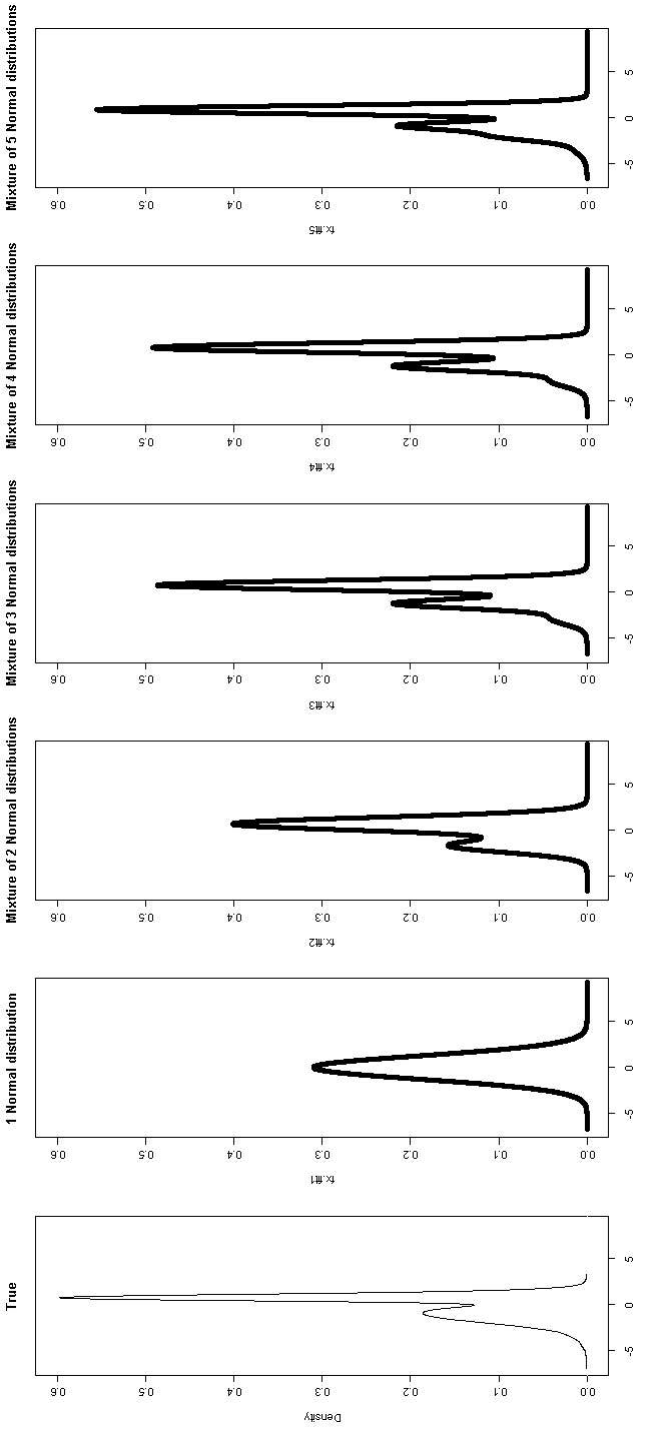

Figure 4.1: Density plots of random effects from simulation results of sensitivity for model-misspecification

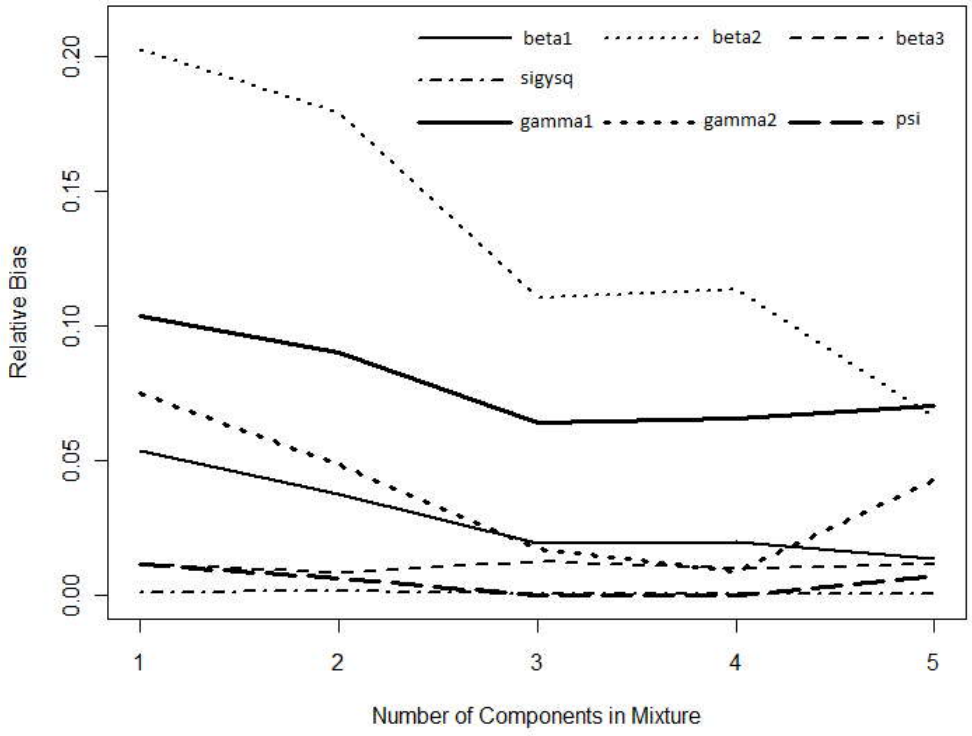

Figure 4.2: Relative bias plot of parameters in longitudinal and hazard models (thin and thick lines respectively) from simulation results of sensitivity for model-misspecification

Table 4.4: Summary of simulation results: Frequencies on the selected number of Normal distributions in mixture (n=200)

	Selected # of Normal distributions in mixture				
Criteria	1	2	3	4	5
AIC	0	0	969	21	10
BIC	0	0	990	5	5

AIC (or BIC), with the one having the lowest AIC (or BIC) being the best. Continuous longitudinal outcomes and survival time are considered with the same setting used in Section 4.5.1. Random effects are generated from a mixture of 3 normal distributions. We fit 5 sets of simultaneous models with different mixtures for random effects which are 1 normal distribution without mixture and the mixtures of 2, 3, 4 and 5 normal distributions. AIC and BIC values are calculated for all 5 fitted mixture models in each data set and we report frequencies of mixture models selected as best by AIC and BIC among 1000 data sets. We consider sample sizes of 200 and 800.

In Table 4.4, we summarize the results for the sample size of 200. We see that both AIC and BIC mostly select the true distribution of a mixture of 3 normal distributions as best. For the large sample size of 800, the mixture of 3 normal distributions is selected by both AIC and BIC for all 1000 simulated data sets. This demonstrates that the number of mixture distributions is properly selected by AIC and BIC even for small sample sizes.

4.6 Analysis of the CHANCE Study

The Carolina Head and Neck Cancer Study (CHANCE) is a population based epidemiologic study conducted at 60 hospitals in 46 counties in North Carolina from 2002 through 2006 (Divaris *et al.* 2010). Patients were diagnosed with head and neck cancer (oral, pharynx, and larynx cancer) from 2002–2006. Their survival status was collected

up to 2007 and QoL was evaluated over time for three years after diagnosis. QoL information was collected through questionnaires. Based on summary scores of the five domains of self-perceived quality of life including Physical Well-Being (PWB), Social/Family Well-Being (SWB), Emotional Well-Being (EWB), Functional Well-Being (FWB) and Head and Neck Cancer Specific symptoms (HNCS), patient's QoL information was classified into satisfaction or dissatisfaction with life. Survival time is defined as the time to death from diagnosis. Demographic and life style characteristics, medical histories and clinical factors are also collected. Ending in December 2009 and excluding the patients with missing data, information on QoL has been obtained from 554 head and neck cancer patients. Based on the death information through 2007 available from the National Death Index (NDI), 85 of 554 patients died and the censoring rate is 85%. The number of observations per patient ranges 1 to 3 with average of 1.93. It is of interest to elucidate the variables which are associated with both QoL satisfaction and survival time for patients with head and neck cancer. In particular, we are interested in the comparison between African-Americans and Whites since it is known that African-Americans have a higher incidence of head and neck cancer and worse survival than Whites. The longitudinal QoL satisfaction outcomes and survival time are correlated within a patient, and this dependency should be taken into account in the analysis.

We apply our proposed method to Head and Neck Cancer Specific symptoms (HNCS) among QoL domains with survival time. Longitudinal HNCS QoL outcomes are binary measurements with 1 ("satisfied") and 0 ("dissatisfied"). We are interested in investigating which factors are related to QoL satisfaction and the risk of death. In the full models for both longitudinal QoL and survival time, we consider race (African-Americans, Whites), the number of 12 oz. beers consumed per week (None, <1, 1–4, 5–14, 15–29, $\geq$ 30), household income (0–10K, 20–30K, 40–50K, $\leq$ 60K), surgery (Yes/No), radiation therapy (Yes/No), chemotherapy (Yes/No), primary tumor site

(Oral & Pharyngeal, Laryngeal) and tumor stage (I, II, III, IV) as categorical, and age at diagnosis (range: 24–80), the number of persons supported by household income (range: 1–5), body mass index (BMI) (range: 15.66–56.28) and the total number of medical conditions reported (range: 0–6) as continuous. Additionally, 2 interactions with race, i.e. race × the total number of medical conditions reported and race × tumor site, are included in both models since we are particularly interested in the difference of QoL and survival between African American and White. Time at survey measurement is also included as a covariate for longitudinal outcomes. A random intercept for the dependence between the QoL satisfaction and the risk of death is included in both models, and assumed to follow a mixture of normal distributions.

For the full model, we first considered 5 different distributions for random effects which are 1 normal distribution without mixture and the mixtures of 2, 3, 4 and 5 normal distributions, and both AIC and BIC selected a mixture of 3 normal distributions with their lowest values as best. Then, we conducted backward variable selection based on the Likelihood Ratio Test (LRT) from the full model assuming the mixture of 3 normal distributions for random effects. Table 4.5 gives the results from the final models after removing non-significant covariates by LRT. From the "Simultaneous" columns, we see the number of 12 oz. beers consumed per week, household income and tumor stage are significantly associated with both patients' HNCS QoL satisfaction and hazard of death. Using 30 or more of 12 oz. beers consumed per week as the reference group, all categories of the smaller amount are associated with higher odds of being satisfied while the categories of 'none' and '5 to 14' of 12 oz. beers consumed per week are associated with lower risk of death. Higher household income is generally associated with higher odds of being satisfied and lower risk of death. Both patients' HNCS QoL satisfaction and risk of death are significantly different for patients in different tumor stages. On the other hand, race (African-American), radiation therapy, the number of

Table 4.5: Results from final models of simultaneous and separate analyses for the Quality of Life and survival time for the CHANCE study

Parameter		Simultaneous Est.	Simultaneous ESE	Simultaneous P-value	Separate Est.	Separate ESE	Separate P-value
HNCS QoL longitudinal model							
Intercept	β_0				1.190	.390	.002
Race (ref= White): African American	β_1	.900	.399	.024	.511	.256	.047
# of 12 oz. beers consumed per week (ref=30 or more)							
– None	β_2	.858	.428	.045	.622	.300	.038
– less than 1	β_3	1.119	.600	.062	.735	.396	.064
– 1 to 4	β_4	1.588	.563	.005	1.268	.326	<.001
– 5 to 14	β_5	1.450	.428	.001	1.018	.279	<.001
– 15 to 29	β_6	1.007	.531	.058	.547	.327	.095
Household income (ref= level1: 0–10K)							
– level2: 20–30K	β_7	- .337	.358	.346	- .328	.258	.204
– level3: 40–50K	β_8	.633	.440	.151	.250	.282	.376
– level4: ≥ 60K	β_9	1.960	.509	<.001	1.045	.286	<.001
Radiation therapy (ref= No) : Yes	β_{10}	-1.668	.608	.006	-1.048	.280	<.001
Tumor stage (ref= I)							
– II	β_{11}	- .683	.554	.218	- .352	.330	.286
– III	β_{12}	-2.012	.534	<.001	-1.198	.314	<.001
– IV	β_{13}	-1.826	.507	<.001	-1.057	.277	<.001
# of persons supported by household income	β_{14}	- .388	.140	.006			
BMI	β_{15}	.061	.026	.021			
Time at survey measurement (years)	β_{16}	.354	.093	<.001	.254	.067	<.001
Hazards model							
Random effect coefficient	ψ	- .206	.078	.008			
# of 12 oz. beers consumed per week (ref=30 or more)							
– None	γ_1	- .705	.347	.042			
– less than 1	γ_2	- .156	.393	.692			
– 1 to 4	γ_3	- .712	.385	.064			
– 5 to 14	γ_4	- .991	.348	.004			
– 15 to 29	γ_5	- .579	.370	.117			
Household income (ref= level1: 0–10K)							
– level2: 20–30K	γ_6	- .206	.274	.453	- .219	.263	.406
– level3: 40–50K	γ_7	- .884	.341	.010	- .928	.331	.005
– level4: ≥ 60K	γ_8	-1.401	.374	<.001	-1.393	.358	<.001
Tumor stage (ref= I)							
– II	γ_9	- .255	.443	.564	- .295	.435	.498
– III	γ_{10}	.168	.403	.677	.136	.389	.727
– IV	γ_{11}	.950	.306	.002	.914	.295	.002
Total # of medical conditions reported	γ_{12}	.207	.095	.030	.205	.091	.025

P-value for testing σ_b^2 being zero is based on a mixture of 0 and χ^2 distribution with 1 degree of freedom with equal mixing probabilities.

persons supported by household income and BMI are selected only in the HNCS QoL longitudinal model while the number of medical conditions reported is significant only in the hazard model. The results indicate that African-Americans, patients not treated with radiation therapy, patients in the family with the smaller number of persons supported by household income, or patients with higher BMI are associated with higher odds of being satisfied, but the risk of death is not affected by these factors. On the other hand, higher number of reported medical conditions is associated with higher risk of death, but it is not associated with HNCS QoL satisfaction. Furthermore, time at survey measurement is statistically significant in the HNCS QoL longitudinal model implying that patients have higher odds to be satisfied over time. The parameter ψ for the dependence between longitudinal HNCS QoL and survival time is negative and is statistically significant with p-value as 0.008. This means the longitudinal HNCS QoL and survival time are correlated and some latent factors which increase HNCS QoL satisfaction also decrease the risk of death. Although not provided in Table 4.5, we have additional parameters of the mixture distribution for random effects in the simultaneous modeling. The obtained estimates of three means of random effects are -3.146, 0.376, and 1.730 with estimated standard errors of 1.284, 2.897 and 0.986 and p-values of 0.014, 0.897 and 0.079, respectively. The first and second mixture components have the weight estimates of 0.147 and 0.105 with estimated standard errors of 0.062 and 0.051 and p-values of 0.018 and 0.037, respectively, and the common variance estimate of random effects is 0.637 with its estimated standard error of 1.286 and p-value of 0.483. In particular, the two weights of mixture components are significant at significant level 0.05, which strengthens the mixture of 3 normal distributions with the estimated 3 means of random effects.

For the purpose of comparison, we also conducted separate analyses for longitudinal HNCS QoL and survival time whose results are given in the last three columns of Table

4.5. Comparing the results from the simultaneous and separate analyses in Table 4.5, we can see our simultaneous analysis identifies two additional factors (the number of persons supported by household income and BMI) in the HNCS QoL longitudinal model and one additional factor (the number of 12 oz. beers consumed per week) in the hazard model.

Figure 4.3 shows the estimated baseline cumulative hazard rates over follow-up time with 95% confidence interval. The estimated baseline cumulative hazard rates look flat at the very early time within a year, but soon appear to be linearly increasing. Figure 4.4 shows the predicted conditional longitudinal trend of HNCS QoL satisfaction probabilities based on the simultaneous models (solid line) and the empirical longitudinal trend of HNCS QoL satisfaction probabilities (dotted line) based on the empirical longitudinal HNCS QoL satisfaction probabilities (dots). The predicted conditional probability of HNCS QoL satisfaction is calculated as the conditional expectation of the conditional probability of HNCS QoL satisfaction given the subject is alive at time t. That is, $\mathrm{E}_{b,\alpha}\left[\mathrm{P}(Y(t) = 1|T > t) \mid \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}\right]$ using model notations in Section 4.2. The empirical probability of HNCS QoL satisfaction is calculated for every 0.05 unit of time at survey measurements. From Figure 4.4, the longitudinal trend of HNCS QoL satisfaction probabilities appears to be increasing over time and the empirical probabilities also gradually increase over time.

4.7 Concluding Remarks

We have relaxed normality assumption of random effects in the simultaneous modeling of longitudinal outcomes and survival time. Assuming the underlying distribution of random effects to be unknown, we used a mixture of Gaussian distributions as an approximation for the random effect distribution. We developed a maximum likelihood estimation method for the proposed simultaneous models and presented asymptotic

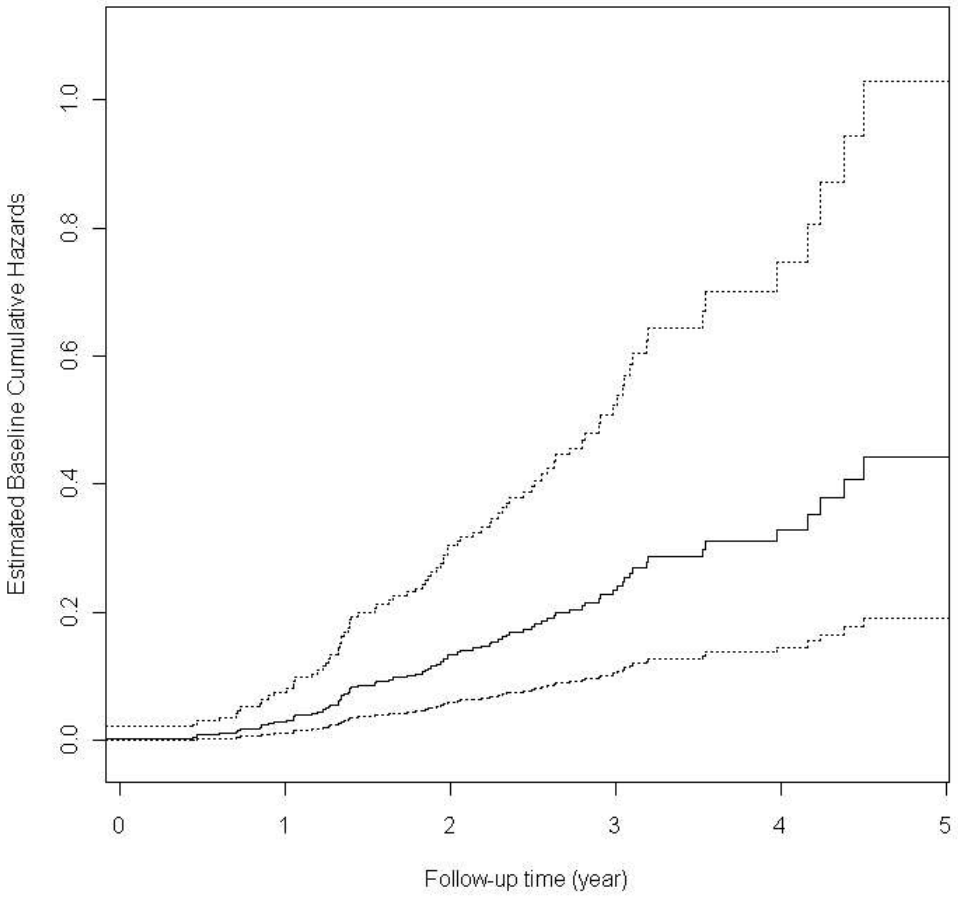

Figure 4.3: Estimated baseline cumulative hazards (solid line) with 95% confidence interval (dotted lines) by the simultaneous analysis of HNCS QoL longitudinal outcome and survival time

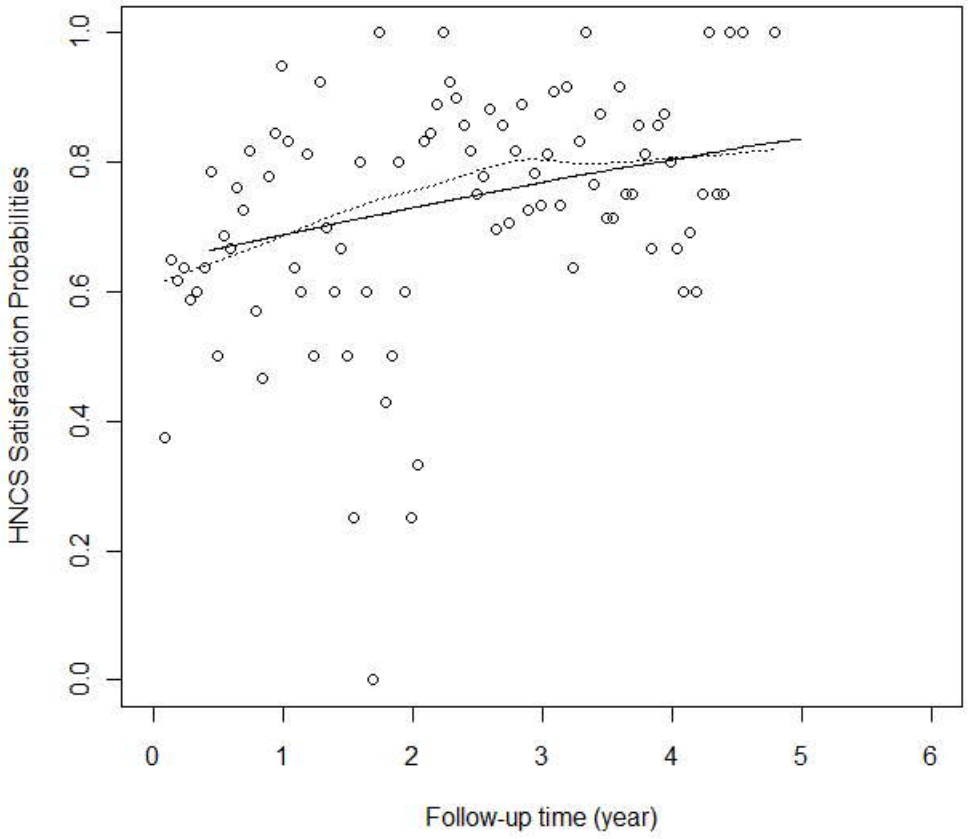

Figure 4.4: The predicted conditional longitudinal trend based on the simultaneous models (solid line) and the empirical longitudinal trend (dotted line) based on the empirical longitudinal HNCS QoL satisfaction probabilities (dots)

properties of the proposed estimators. The proposed estimation procedure using EM algorithm has been assessed via simulation studies for both continuous and binary longitudinal data with survival time. The proposed estimates performed well in finite samples. The variance estimates based on the observed information matrix approximate the true variance well in finite samples. Simulation studies indicated that, when the true distribution of random effects is not normal, mixture distributions yield less biased estimates than no mixture and all the estimated density plots of random effects based on mixture distributions appear to have similar shapes to the true distribution. Furthermore, simulation studies also showed that the number of mixture distributions is properly selected by AIC and BIC. The proposed method was applied to data from the CHANCE study.

Our proposed method is an effort to relax the assumption that the random effects come from a normal distribution which is often made for computational reasons. This normality assumption is difficult to check because random effects are latent and never observed. Furthermore, if this assumption fails to hold, the estimates of the parameters in the generalized linear mixed model and the hazards model are biased. In this paper, the mixture of normal distributions has been shown to be a good approximation for the random effects in the simultaneous modeling when the underlying distribution of random effects is unknown. The advantage of this approach is that many continuous distributions can be well approximated by a finite normal mixture which implies that our proposed method will generally perform well. When sample size and the number of observations per subject are too large, computation may be intensive due to the integration of complete data likelihood over random effects. It will be of interest to develop a more computationally efficient approach. One possibility is to consider a penalized likelihood approach by the Laplace approximation which is currently under investigation by us.

Chapter 5

PENALIZED LIKELIHOOD APPROACH FOR JOINT ANALYSIS OF SURVIVAL TIME AND LONGITUDINAL OUTCOMES

5.1 Introduction

In biomedical or public health research, it is common that both longitudinal outcomes over time and survival endpoint are collected for the same subject along with the subject's characteristics or risk factors. Investigators are interested in finding important variables which predict both longitudinal outcomes and survival time. Among the existing approaches for longitudinal data and survival time, the selection model and the pattern mixture model have been widely used. The selection model estimating the distribution of survival time given longitudinal data was studied by numerous authors, for example, Tsiatis, Degruttola, and Wulfsohn (1995), Tsiatis and Davidian (2001), Xu and Zeger (2001a,b) and Tseng, Hsieh and Wang (2005). The pattern mixture model focuses on the trend of longitudinal outcomes conditional on survival time and was studied by Wu and Carroll (1988), Hogan and Laird (1997), Albert and Follmann (2000, 2007) and Ding and Wang (2008) among others. On the other hand, simultaneous

modeling of the longitudinal and survival data was proposed by Xu and Zeger (2001b), Zeng and Cai (2005), Elashfoff, Li and Ni (2007, 2008) and Rizopoulos, Verbeke, Lesaffre and Vanrenterghem (2008). Wang and Taylor (2001), Brown and Ibrahim (2003) and Hu, Li and Li (2009) studied simultaneous modeling in the Bayesian perspective.

In the joint models, random effects are incorporated to accommodate the latent dependence between survival time and longitudinal outcomes, and often assumed to be normally distributed so that we can integrate a complete data likelihood over random effects to obtain a full likelihood. The maximum likelihood approach using an Expectation-Maximization algorithm provides the estimators which are asymptotically consistent and follows an asymptotic Gaussian process. However, the EM algorithm may be intensive on computation with large sample sizes and large numbers of longitudinal observations per subject. In the view of the cumbersome and often intractable numerical integrations required for a full likelihood, one possible alternative can be the penalized likelihood approach which gives a penalty for regarding random effects as fixed effects in the likelihood obtained by Laplace approximation. In generalized linear mixed models (GLMM), the penalized quasi-likelihood (PQL) approach is the most common estimation procedure. The PQL was proposed as an approximate Bayes procedure for some commonly occurring GLMM's by Laird (1978) and the PQL method exploited by Green (1987) for semiparametric regression analysis is available for inference in hierarchical models where the focus is on shrinkage estimation of the random effects (Robinson, 1991). Breslow and Clayton (1993) proposed to use the PQL with some modifications to a Laplace expansion for a GLMM in order to motivate standard estimating equations that may be solved by iterative application of normal theory procedures. Breslow and Lin (1995) and Lin and Brelsow (1996) derived the general expressions for the asymptotic biases in approximate estimators of regression coefficients and variance component in the GLMMs with a single source of extraneous variation

and multiple components of dispersion, respectively. The PQL also has been studied in a wide variety of GLMMs by Bartlett and Sutradhar (1999), Huber and Victoria-Feser (2004), Localio, Berlin and Ten Have (2006), Nelson and Leroux (2008), Dang, Mazumdar and Houck (2008), Jang and Lim (2009), and Masaoud and Stryhn (2010). Furthermore, the PQL is already built in SAS GLIMMIX procedure and used for the analysis of the GLMM. On the other hand, Ripatti and Palmgren (2000) proposed a penalized partial likelihood for multivariate frailty models in survival analysis. In joint modeling framework, Ye, Lin and Taylor (2008) proposed a penalized joint likelihood for a selection model and considered a continuous longitudinal process to be included as a covariate for survival time. Their penalized joint likelihood is obtained by replacing the full survival likelihood with a partial likelihood in the Laplace approximation to the full joint likelihood function, which is not equal to the actual form derived from the full joint likelihood function. On the other hand, there is no work done on the penalized likelihood approach for the simultaneous modeling of longitudinal outcomes and survival time. Furthermore, the previous study using the penalized likelihood in joint analysis (Ye, Lin and Taylor, 2008) considered continuous longitudinal data from a normal distribution.

In this paper, we propose to use a penalized likelihood to develop a more efficient estimation procedure on computation for simultaneous modeling than the EM algorithm of the maximum likelihood approach. We consider a generalized linear mixed model for longitudinal outcomes to incorporate both categorical and continuous data and a stratified Cox proportional hazards model for survival time. In this estimation procedure, all the parameters are estimated together at the same time. If the EM algorithm of maximum likelihood approach performs similarly to the penalized likelihood method on computational time. it will be better to use the full likelihood. In the meantime, if the penalized likelihood method takes less time and provides unbiased and consistent

estimates similar to those from EM algorithm, the penalized likelihood method will be preferred.

The organization of this paper is as follows. We present a simultaneous modeling for longitudinal outcomes and survival time with random effects in Section 5.2 and describe the proposed estimation procedure in Section 5.3. Numerical results from simulation studies are given in Section 5.4, and our proposed method is illustrated with the data from the Carolina Head and Neck Cancer Study (CHANCE) in Section 5.5. In Section 5.6, we discuss some further consideration.

5.2 Model Formulation and Notation

We use $Y(t)$ to denote the value of a longitudinal marker process at time t. Suppose $Y(t)$ is from a distribution belonging to exponential family in order to incorporate both continuous and categorical measurements. Let T denote survival time, and suppose that the survival time T is possibly right censored. Suppose a set of n subjects are followed over an interval $[0, \tau]$, where τ is the study end time. Denote $\boldsymbol{b}_i$, $i = 1, \ldots, n$, as a vector of subject-specific random effects of dimension d_b and $\boldsymbol{b}_i$'s are mutually independent and identically distributed from a multivariate normal with mean zero and covariance matrix $\boldsymbol{\Sigma}_b$.

Given the random effects $\boldsymbol{b}_i$, the observed covariates, and the observed outcome history till time t, we assume that the longitudinal outcome $Y_i(t)$ at time t for subject i follows a distribution from the exponential family with density,

$$\exp\left\{\frac{y_i\eta_i(t) - B(\eta_i(t))}{A(D_i(t;\phi))} + C(y_i, D_i(t;\phi))\right\} \tag{5.1}$$

with $\mu_i(t) = \mathrm{E}(Y_i(t)|\boldsymbol{b}_i) = B'(\eta_i(t))$ and $v_i(t) = \mathrm{Var}(Y_i(t)|\boldsymbol{b}_i) = B''(\eta_i(t))A(D_i(t;\phi))$, satisfying

$$\eta_i(t) = g(\mu_i(t)) = \boldsymbol{X}_i(t)\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_i(t)\boldsymbol{b}_i$$

and $v_i(t) = v(\mu_i(t))A(D_i(t;\phi))$, where $g(\cdot)$ and $v(\cdot)$ are known link and variance functions respectively, $\boldsymbol{X}_i(t)$ and $\widetilde{\boldsymbol{X}}_i(t)$ are the row vectors of the observed covariates for subject i, and $\boldsymbol{\beta}$ is a column vector of coefficients for $\boldsymbol{X}_i(t)$. The random effect $\boldsymbol{b}_i$ is allowed to differ for different individuals. Additionally, $\boldsymbol{X}_i(t)$ and $\widetilde{\boldsymbol{X}}_i(t)$ can be completely different or share some components, and may include dummy variables for different strata.

Given the random effects $\boldsymbol{b}_i$, the observed covariates, and the observed survival history before time t, the conditional hazard rate function for the survival time T_i of subject i is assumed to follow a stratified multiplicative hazards model,

$$\lambda_s(t)\exp\{\widetilde{\boldsymbol{Z}}_i(t)(\boldsymbol{\psi}\circ\boldsymbol{b}_i) + \boldsymbol{Z}_i(t)\boldsymbol{\gamma}\}, \tag{5.2}$$

where $\boldsymbol{Z}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ are the row vectors of the observed covariates and may share some components, $\boldsymbol{\psi}$ is a vector of parameters of the coefficients for random effects, $\lambda_s(t)$ is the s-th stratum baseline hazard rate function, and $\boldsymbol{\gamma}$ is a column vector of coefficients for $\boldsymbol{Z}_i(t)$. Note that $\boldsymbol{Z}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ do not include dummy variables for strata since baseline hazard rate is stratum-specific. Here, for any vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ of the same dimension, $\boldsymbol{a}_1\circ\boldsymbol{a}_2$ denotes the component-wise product. In addition, $\widetilde{\boldsymbol{X}}_i(t)$ and $\widetilde{\boldsymbol{Z}}_i(t)$ have the same dimensions as $\boldsymbol{b}_i$'s.

Under models (5.1) and (5.2), the two outcomes $Y(t)$ and T are independent conditional on the covariates and random effect. The parameter $\boldsymbol{\psi}$ in model (5.2) characterizes the dependence between the longitudinal outcomes and the survival time due to latent random effect: $\boldsymbol{\psi} = \mathbf{0}$ means that the dependence between the survival time and longitudinal responses are not due to these latent variables; $\boldsymbol{\psi} \neq \mathbf{0}$ means that such de-

pendence may be due to these latent variables. In other words, $\boldsymbol{\psi} > \mathbf{0}$ implies that there may be some latent factors increasing both the longitudinal outcomes and the risk of survival endpoint simultaneously while $\boldsymbol{\psi} < \mathbf{0}$ implies that some latent factors causing the increment of longitudinal outcomes may decrease the risk of survival endpoint.

Let n_i be the number of the observed longitudinal measurements for subject i, and assume that the distributions of n_i and the observation times for longitudinal measurements are independent of the parameters of interest in this joint model. The observed data from n subjects are $(n_i, Y_{ij}, \boldsymbol{X}_{ij}, \widetilde{\boldsymbol{X}}_{ij})$, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, and $(V_i, \Delta_i, S_i, \{(\boldsymbol{Z}_i(t), \widetilde{\boldsymbol{Z}}_i(t)) : t \leq V_i\})$, $i = 1, \ldots, n$, where for subject i, $(Y_{ij}, \boldsymbol{X}_{ij}, \widetilde{\boldsymbol{X}}_{ij})$ is the j-th observation of $(Y_i(t), \boldsymbol{X}_i(t), \widetilde{\boldsymbol{X}}_i(t))$, C_i is the right-censoring time and independent of T_i and $Y_i(t)$ given the covariates and the random effects, $V_i = \min(T_i, C_i)$, S_i denotes the stratum, and $\Delta_i = I(T_i \leq C_i)$.

Our goal is to estimate and make inferences on the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\phi}^T, \mathrm{Vec}(\boldsymbol{\Sigma}_b)^T, \boldsymbol{\psi}^T, \boldsymbol{\gamma}^T)^T$ and the baseline cumulative hazard functions with S strata, $\boldsymbol{\Lambda}(t) = (\Lambda_1(t), \ldots, \Lambda_S(t))^T$, where $\Lambda_s(t) = \int_0^t \lambda_s(u)du$, $s = 1, \ldots, S$. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are from the longitudinal model, $\boldsymbol{\psi}$ and $\boldsymbol{\gamma}$ are from the hazard model, and $\boldsymbol{\Sigma}_b$ is associated with the random effects. $\mathrm{Vec}(\cdot)$ operator creates a column vector from a matrix by stacking the diagonal and upper-triangle elements of the matrix.

5.3 Estimation Procedure

For all n subjects, we write $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T$, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, $\boldsymbol{V} = (V_1, \ldots, V_n)^T$, and $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T)^T$. Then, the likelihood function of the complete data $(\boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ has the form,

$$
\begin{aligned}
& L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) \\
&= \prod_{s=1}^{S} \prod_{i=1}^{n} \left[f(\boldsymbol{Y}_i, V_i | \boldsymbol{b}_i) f(\boldsymbol{b}_i) \right]^{I(S_i=s)} = \prod_{i=1}^{n} f(\boldsymbol{Y}_i | \boldsymbol{b}_i) \left(\prod_{s=1}^{S} \left[f(V_i | \boldsymbol{b}_i) \right]^{I(S_i=s)} \right) f(\boldsymbol{b}_i) \\
&= \prod_{i=1}^{n} \exp \left\{ \sum_{j=1}^{n_i} \left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i) - B(\boldsymbol{\beta}; \boldsymbol{b}_i)}{A(D_i(t_j;\phi))} + C(Y_{ij}; D_i(t_j;\phi)) \right] \right\} \\
&\quad \times \left(\prod_{s=1}^{S} \left[\lambda_s(V_i)^{\Delta_i} \exp \left\{ \Delta_i \left[\widetilde{\boldsymbol{W}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{W}_i(V_i)\boldsymbol{\gamma} \right] \right. \right. \right. \\
&\qquad \left. \left. \left. - \int_0^{V_i} \exp \left\{ \widetilde{\boldsymbol{W}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{W}_i(u)\boldsymbol{\gamma} \right\} d\Lambda_s(u) \right\} \right]^{I(S_i=s)} \right) \\
&\quad \times (2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \exp \left\{ -\frac{1}{2} \boldsymbol{b}_i^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}_i \right\},
\end{aligned}
$$

and the full likelihood function of the observed data $(\boldsymbol{Y}, \boldsymbol{V})$ for the parameter $(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ is expressed as

$$
L_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) = \int_{\boldsymbol{b}} L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) d\boldsymbol{b}. \tag{5.3}
$$

The primary difficulty in implementing this full likelihood inference lies in the integrations needed to evaluate the complete data likelihood $L_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ and its partial derivatives.

In the EM algorithm of maximum likelihood approach, the random effect $\boldsymbol{b}_i$ is considered as missing data for $i = 1, \ldots, n$. Thus, the M-step solves the conditional score equations from complete data log-likelihood given observations, where the conditional expectation is evaluated in the E-step. The procedure involves iterating between the two steps until convergence is achieved. In the E-step calculating the conditional expectations of some known functions of $\boldsymbol{b}_i$ needed in the next M-step, a numerical approximation method such as the Gauss-Hermite Quadrature is required for the integration with the posterior probability of random effects. When sample size (n), the

number of observations per subject (n_i), and the number of parameters to be estimated are large, the task involving the integration in the E-step is very challenging. Therefore, to make the simultaneous modeling more practical, we build the algorithm which relieves the computational burden.

Our proposed estimation method is to calculate the maximum penalized likelihood estimates for $(\boldsymbol{\theta}, \boldsymbol{\Lambda}(t))$ over a set in which $\boldsymbol{\theta}$ is in a bounded set and $\Lambda_s(t)$ of $\boldsymbol{\Lambda}(t)$ belongs to a space consisting of all the increasing functions with $\Lambda_s(0) = 0$, $s = 1, \ldots, S$. We let each $\Lambda_s(t)$ of $\boldsymbol{\Lambda}(t)$, $s = 1, \ldots, S$, be an increasing and right-continuous step function with jumps only at the observed failure times belonging to stratum s. The penalized likelihood is obtained by Laplace approximation, and the proposed approach is expected to be less intensive in computation in the sense that it imposes the penalty for considering the random effect as the fixed effect in the likelihood and therefore no calculation for integrating the likelihood over random effects is needed.

5.3.1 Laplace approximation

The full likelihood (5.3) can be written as

$$L_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) = (2\pi)^{-nd_b/2} |\boldsymbol{\Sigma}_b|^{-n/2} \int_{\boldsymbol{b}} \exp\left\{ \sum_{i=1}^{n} \left[l_{i|b}(\boldsymbol{\theta}, \Lambda_s) - \frac{1}{2} \boldsymbol{b}_i^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}_i \right] \right\} d\boldsymbol{b}, \quad (5.4)$$

where the logarithm of the conditional joint density given an unobserved random effect $\boldsymbol{b}_i$ is

$$\begin{aligned} l_{i|b}(\boldsymbol{\theta}, \Lambda_s) = & \sum_{j=1}^{n_i} \left[\frac{Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i) - B_{ij}(\boldsymbol{\beta}; \boldsymbol{b}_i)}{A(D_i(t_j; \phi))} + C(Y_{ij}; D_i(t_j; \phi)) \right] \\ & + \sum_{s=1}^{S} I(S_i = s) \Bigg[\Delta_i \log(\lambda_s(V_i)) + \Delta_i [\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma}] \\ & - \int_0^{V_i} \exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma}\} d\Lambda_s(u) \Bigg]. \quad (5.5) \end{aligned}$$

Then, we have the following form of the full log-likelihood,

$$l_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) = \sum_{i=1}^{n} \left[-\frac{d_b}{2}\log(2\pi) - \frac{1}{2}|\boldsymbol{\Sigma}_b| + l_{i|b}(\boldsymbol{\theta}, \Lambda_s) - \frac{1}{2}\boldsymbol{b}_i^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}_i \right]. \tag{5.6}$$

In (5.4), define

$$-\boldsymbol{\kappa}(\boldsymbol{b}) \;=\; \sum_{i=1}^{n} \left[l_{i|b}(\boldsymbol{\theta}, \Lambda_s) - \frac{1}{2}\boldsymbol{b}_i^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}_i \right] = \sum_{i=1}^{n} \left[-\boldsymbol{\kappa}_i(\boldsymbol{b}_i) \right] \tag{5.7}$$

and apply Laplace's approximation as following,

$$-\boldsymbol{\kappa}_i(\boldsymbol{b}_i) \;\approx\; -\boldsymbol{\kappa}_i(\widetilde{\boldsymbol{b}}_i) - \frac{1}{2}(\boldsymbol{b} - \widetilde{\boldsymbol{b}}_i)^T \boldsymbol{\kappa}_i''(\widetilde{\boldsymbol{b}}_i)(\boldsymbol{b}_i - \widetilde{\boldsymbol{b}}_i),$$

where $\boldsymbol{\kappa}'$ and $\boldsymbol{\kappa}''$ denote the d_b vector and $d_b \times d_b$ dimensional matrix of first- and second-order partial derivatives of $\boldsymbol{\kappa}$ with respect to $\boldsymbol{b}$ and $\widetilde{\boldsymbol{b}}$ denotes the solution to $\boldsymbol{\kappa}'(\boldsymbol{b}) = 0$ that minimizes $\boldsymbol{\kappa}(\boldsymbol{b})$. Then, the full likelihood function (5.4) can be approximated as followings,

$$\begin{aligned}
& L_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \boldsymbol{Y}, \boldsymbol{V}) \\
= \;& (2\pi)^{-nd_b/2} |\boldsymbol{\Sigma}_b|^{-n/2} \int_{\boldsymbol{b}} \exp\left\{ -\boldsymbol{\kappa}(\boldsymbol{b}) \right\} d\boldsymbol{b} = \prod_{i=1}^{n} \left[(2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \int_{\boldsymbol{b}} \exp\left\{ -\boldsymbol{\kappa}_i(\boldsymbol{b}_i) \right\} d\boldsymbol{b} \right] \\
\approx \;& \prod_{i=1}^{n} \left[(2\pi)^{-d_b/2} |\boldsymbol{\Sigma}_b|^{-1/2} \int_{\boldsymbol{b}} \exp\left\{ -\boldsymbol{\kappa}_i(\widetilde{\boldsymbol{b}}_i) - \frac{1}{2}(\boldsymbol{b}_i - \widetilde{\boldsymbol{b}}_i)^T \boldsymbol{\kappa}_i''(\widetilde{\boldsymbol{b}}_i)(\boldsymbol{b}_i - \widetilde{\boldsymbol{b}}_i) \right\} d\boldsymbol{b} \right] \\
= \;& \prod_{i=1}^{n} \left[|\boldsymbol{\Sigma}_b|^{-1/2} \exp\left\{ -\boldsymbol{\kappa}_i(\widetilde{\boldsymbol{b}}_i) \right\} |\boldsymbol{\kappa}_i''(\widetilde{\boldsymbol{b}}_i)|^{-1/2} \right. \\
& \qquad \left. (2\pi)^{-d_b/2} |\boldsymbol{\kappa}_i''(\widetilde{\boldsymbol{b}}_i)|^{1/2} \int_{\boldsymbol{b}} \exp\left\{ -\frac{1}{2}(\boldsymbol{b}_i - \widetilde{\boldsymbol{b}}_i)^T \boldsymbol{\kappa}_i''(\widetilde{\boldsymbol{b}}_i)(\boldsymbol{b}_i - \widetilde{\boldsymbol{b}}_i) \right\} d\boldsymbol{b} \right] \\
= \;& \prod_{i=1}^{n} \left[|\boldsymbol{\Sigma}_b|^{-1/2} \exp\left\{ -\boldsymbol{\kappa}_i(\widetilde{\boldsymbol{b}}_i) \right\} |\boldsymbol{\kappa}_i''(\widetilde{\boldsymbol{b}}_i)|^{-1/2} \right] \\
= \;& |\boldsymbol{\Sigma}_b|^{-n/2} \exp\left\{ \sum_{i=1}^{n} \left[-\boldsymbol{\kappa}_i(\widetilde{\boldsymbol{b}}_i) - \frac{1}{2}\log|\boldsymbol{\kappa}_i''(\widetilde{\boldsymbol{b}}_i)| \right] \right\}.
\end{aligned} \tag{5.8}$$

In the above (5.8), ignoring the multiplicative constant, the approximation yields

$$-\boldsymbol{\kappa}(\boldsymbol{b}) \approx -\frac{1}{2}\log|\boldsymbol{\kappa}''(\widetilde{\boldsymbol{b}})| - \boldsymbol{\kappa}(\widetilde{\boldsymbol{b}}).$$

Note that, from (5.7),

$$\begin{aligned}
\boldsymbol{\kappa}_i(\widetilde{\boldsymbol{b}}_i) &= -\widetilde{l}_{i|b}(\boldsymbol{\theta}, \Lambda_s) + \frac{1}{2}\widetilde{\boldsymbol{b}}_i^T \boldsymbol{\Sigma}_b^{-1}\widetilde{\boldsymbol{b}}_i, \\
\boldsymbol{\kappa}_i'(\widetilde{\boldsymbol{b}}_i) &= -\widetilde{l}'_{i|b}(\boldsymbol{\theta}, \Lambda_s) + \boldsymbol{\Sigma}_b^{-1}\widetilde{\boldsymbol{b}}_i, \\
\boldsymbol{\kappa}_i''(\widetilde{\boldsymbol{b}}_i) &= -\widetilde{l}''_{i|b}(\boldsymbol{\theta}, \Lambda_s) + \boldsymbol{\Sigma}_b^{-1}, \qquad (5.9)
\end{aligned}$$

where $\widetilde{l}_{i|b}(\boldsymbol{\theta}, \Lambda_s)$ is (5.5) evaluated at $\widetilde{\boldsymbol{b}}_i$, and $\widetilde{l}'_{i|b}(\boldsymbol{\theta}, \Lambda_s)$ and $\widetilde{l}''_{i|b}(\boldsymbol{\theta}, \Lambda_s)$ are the first and second derivatives of (5.5) with respect to $\boldsymbol{b}_i$ evaluated at $\widetilde{\boldsymbol{b}}_i$. Then, the first order Laplace approximation to the full likelihood becomes

$$\begin{aligned}
(5.8) &= \exp\left\{\sum_{i=1}^{n}\left[-\frac{1}{2}\log|\boldsymbol{\Sigma}_b| - \boldsymbol{\kappa}_i(\widetilde{\boldsymbol{b}}_i) - \frac{1}{2}\log|-\widetilde{l}''_{i|b}(\boldsymbol{\theta}, \Lambda_s) + \boldsymbol{\Sigma}_b^{-1}|\right]\right\} \\
&= \exp\left\{\sum_{i=1}^{n}\left[-\frac{1}{2}\log|\boldsymbol{I}_{d_b} - \boldsymbol{\Sigma}_b\widetilde{l}''_{i|b}(\boldsymbol{\theta}, \Lambda_s)| + \widetilde{l}_{i|b}(\boldsymbol{\theta}\Lambda_s) - \frac{1}{2}\widetilde{\boldsymbol{b}}_i^T\boldsymbol{\Sigma}_b^{-1}\widetilde{\boldsymbol{b}}_i\right]\right\} \\
&= \exp\left\{\sum_{i=1}^{n}\widetilde{l}_{fi}(\boldsymbol{\theta}, \Lambda_s)\right\} = \exp\left\{\widetilde{l}_f(\boldsymbol{\theta}, \Lambda_s)\right\},
\end{aligned}$$

where $\widetilde{l}_f(\boldsymbol{\theta}, \Lambda_s)$ is the first order Laplace approximation to the full log-likelihood function $l_f(\boldsymbol{\theta}, \Lambda_s)$ of (5.6). That is,

$$\tilde{l}_f(\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \sum_{i=1}^{n}\left[-\frac{1}{2}\log|\boldsymbol{I}_{d_b} - \boldsymbol{\Sigma}_b\widetilde{l}''_{i|b}(\boldsymbol{\theta}, \Lambda_s)| + \left(\widetilde{l}_{i|b}(\boldsymbol{\theta}, \Lambda_s) - \frac{1}{2}\widetilde{\boldsymbol{b}}_i^T\boldsymbol{\Sigma}_b^{-1}\widetilde{\boldsymbol{b}}_i\right)\right], \quad (5.10)$$

where $\tilde{l}''_{i|b}(\boldsymbol{\theta}, \Lambda_s) = l''_{i|b}(\boldsymbol{\theta}, \Lambda_s; \widetilde{\boldsymbol{b}}_i)$ and $\tilde{l}_{i|b}(\boldsymbol{\theta}, \Lambda_s) = l_{i|b}(\boldsymbol{\theta}, \Lambda_s; \widetilde{\boldsymbol{b}}_i)$, and the first and second derivatives of $l_{i|b}(\boldsymbol{\theta}, \Lambda_s; \widetilde{\boldsymbol{b}}_i)$ in (5.5) with respect to $\boldsymbol{b}_i$ are

$$l'_{i|b}(\boldsymbol{\theta}, \Lambda_s) = \sum_{j=1}^{n_i}\left[\frac{Y_{ij}\widetilde{\boldsymbol{X}}_{ij}}{A(D_i(t_j;\phi))} - \frac{B'_{ij}(\boldsymbol{\beta};\boldsymbol{b}_i)}{A(D_i(t_j;\phi))}\right] + \sum_{s=1}^{S} I(S_i=s)\Bigg[\Delta_i(\widetilde{\boldsymbol{Z}}_i(V_i)\circ\boldsymbol{\psi}^T) - \int_0^{V_i}\exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}(\widetilde{\boldsymbol{Z}}_i(u)\circ\boldsymbol{\psi}^T)d\Lambda_s(u)\Bigg]$$

and

$$l''_{i|b}(\boldsymbol{\theta}, \Lambda_s) = -\sum_{j=1}^{n_i}\frac{B''_{ij}(\boldsymbol{\beta};\boldsymbol{b}_i)}{A(D_i(t_j;\phi))} - \sum_{s=1}^{S} I(S_i=s)\int_0^{V_i}\exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_i)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}(\widetilde{\boldsymbol{Z}}_i^T(u)\circ\boldsymbol{\psi})^{\otimes 2}d\Lambda_s(u), \quad (5.11)$$

where $B'_{ij}(\boldsymbol{\beta};\boldsymbol{b}_i)$ and $B''_{ij}(\boldsymbol{\beta};\boldsymbol{b}_i)$ are the first and second derivatives of $B_{ij}(\boldsymbol{\beta};\boldsymbol{b}_i)$ with respect to $\boldsymbol{b}_i$.

5.3.2 Penalized likelihood

Now, we further approximate (5.11). The first term of (5.11) can be expressed as $\widetilde{\boldsymbol{X}}_i^T\boldsymbol{W}_i\widetilde{\boldsymbol{X}}_i$, where $\boldsymbol{W}_i$ is the $n_i \times n_i$ diagonal matrix with $w_{ij} = A(D_i(t_j;\phi))\big[g'(\mu_{ij}^b)\big]^{-1}$, $g(\cdot)$ is a canonical link function, $\mu_{ij}^b = \mathrm{E}\left(Y_{ij}|\boldsymbol{b}_i\right)$, $g'(\mu_{ij}^b)$ is the derivative of $g(\mu_{ij}^b)$ with respect to μ_{ij}^b, and $\widetilde{\boldsymbol{X}}_i = (\widetilde{\boldsymbol{X}}_{i1}^T, \ldots, \widetilde{\boldsymbol{X}}_{in_i}^T)^T$. Generalized linear model (GLM) iterative weights $\boldsymbol{W}_i$ (i.e. w_{ij}) vary slowly or not at all at the function of the mean, and hance, by taking an expectation,

$$\mathrm{E}\left[-\sum_{j=1}^{n_i}\frac{B''_{ij}(\boldsymbol{\beta};\boldsymbol{b}_i)}{A(D_i(t_j;\phi))}\right] = \mathrm{E}\left[\widetilde{\boldsymbol{X}}_i^T\boldsymbol{W}_i\widetilde{\boldsymbol{X}}_i\right] \quad (5.12)$$

becomes a constant. In the second term of (5.11), by taking an expectation, we have

$$\mathrm{E}\left[\int_0^{V_i}\exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}(\widetilde{\boldsymbol{Z}}_i^T(u)\circ\boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(u)\circ\boldsymbol{\psi}^T)d\Lambda_s(u)\right]$$

$$
\begin{aligned}
&= \mathrm{E}\left[\int (\widetilde{\boldsymbol{Z}}_i^T(u) \circ \boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(u) \circ \boldsymbol{\psi}^T) I(V_i \geq u) \exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi} \circ \boldsymbol{b}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma}\} \lambda_s(u) du \right] \\
&= \mathrm{E}\left[\int (\widetilde{\boldsymbol{Z}}_i^T(u) \circ \boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(u) \circ \boldsymbol{\psi}^T) I(T_i \geq u) I(C_i \geq u) \frac{1}{S_{sT}(u)} dF_{sT}(u) \right] \\
&= \mathrm{E}\left[(\widetilde{\boldsymbol{Z}}_i^T(V_i) \circ \boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(V_i) \circ \boldsymbol{\psi}^T) \Delta_i \right]
\end{aligned}
$$

becomes a constant. Thus, the expectation of second term in (5.11)

$$
\sum_{s=1}^{S} I(S_i = s) \, \mathrm{E}\left[- (\widetilde{\boldsymbol{Z}}_i^T(V_i) \circ \boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(V_i) \circ \boldsymbol{\psi}^T) \Delta_i \right] \tag{5.13}
$$

also becomes a constant including $\boldsymbol{\psi}$. Since the expected value of (5.11), $\mathrm{E}\left[l''_{i|b}(\boldsymbol{\theta}, \Lambda_s)\right]$, is the sum of (5.12) and (5.13), $l''_{i|b}(\boldsymbol{\theta}, \Lambda_s)$ in (5.11) also becomes asymptotically a constant. Therefore, in (5.10), $|\boldsymbol{I}_{d_b} - \boldsymbol{\Sigma}_b \widetilde{l''}_{i|b}(\boldsymbol{\theta}, \Lambda_s)|$ is asymptotically a constant including $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\psi}$. Then, we derive the penalized log-likelihood as following,

$$
\begin{aligned}
& l_P(\boldsymbol{\theta}, \boldsymbol{\Lambda}) \\
&= \sum_{i=1}^{n} \Bigg[-\frac{1}{2} \log \left| \boldsymbol{I}_{d_b} - \boldsymbol{\Sigma}_b \left(\mathrm{E}\left[\widetilde{\boldsymbol{X}}_i^T \boldsymbol{W}_i \widetilde{\boldsymbol{X}}_i \right] - \sum_{s=1}^{S} I(S_i = s) \, \mathrm{E}\left[(\widetilde{\boldsymbol{Z}}_i^T(V_i) \circ \boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(V_i) \circ \boldsymbol{\psi}^T) \Delta_i \right] \right) \right| \\
&\qquad + \left(\widetilde{l}_{i|b}(\boldsymbol{\theta}, \Lambda_s) - \frac{1}{2} \widetilde{\boldsymbol{b}}_i^T \boldsymbol{\Sigma}_b^{-1} \widetilde{\boldsymbol{b}}_i \right) \Bigg].
\end{aligned} \tag{5.14}
$$

Since $|\boldsymbol{I}_{d_b} - \boldsymbol{\Sigma}_b \widetilde{l''}_{i|b}(\boldsymbol{\theta}, \Lambda_s)|$ in (5.10) which corresponds to the first term in (5.14) is asymptotically a constant including $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\psi}$, it only contributes to the estimating equations of $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\psi}$, and we ignore the term to obtain the estimating equations of $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$. $-\frac{1}{2} \widetilde{\boldsymbol{b}}_i^T \boldsymbol{\Sigma}_b^{-1} \widetilde{\boldsymbol{b}}_i$ is the penalty term for regarding the random effects as fixed effects by replacing $\boldsymbol{b}$ with $\widetilde{\boldsymbol{b}}$ in the likelihood. We choose $\boldsymbol{\theta}$ to maximize the penalized likelihood $l_p(\boldsymbol{\theta}, \boldsymbol{\Lambda})$ in (5.14). That is, $(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{b}})$ jointly maximize the equation (5.14). The score equations for $(\boldsymbol{\theta}, \boldsymbol{b})$ are obtained by differentiating (5.14) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{b}$, respectively.

5.3.3 Implementation

We conduct the Newton-Rapshon method for estimating equations to obtain $\widetilde{\boldsymbol{b}}$ and $\widehat{\boldsymbol{\theta}}$. The procedure involves iterating between the following two steps until convergence is achieved: at the k-th iteration,

Step1 : Conduct one-step Newton-Rapshon iteration to obtain the solution $\widetilde{\boldsymbol{b}}$ of $\boldsymbol{\kappa}'(\boldsymbol{b}) = 0$. The $(k+1)$-th estimate is $\widetilde{\boldsymbol{b}}^{(k+1)} = \widetilde{\boldsymbol{b}}^{(k)} - \big[\boldsymbol{\kappa}''(\widetilde{\boldsymbol{b}}^{(k)})\big]^{-1}\big[\boldsymbol{\kappa}'(\widetilde{\boldsymbol{b}}^{(k)})\big]^T$, where $\widetilde{\boldsymbol{b}}^{(k)} = \widetilde{\boldsymbol{b}}^{(k)}(\widehat{\boldsymbol{\theta}}^{(k-1)})$, $\boldsymbol{\kappa}'(\boldsymbol{b}) = (\boldsymbol{\kappa}_1'(\boldsymbol{b}_1)^T, \ldots, \boldsymbol{\kappa}_n'(\boldsymbol{b}_n)^T)^T$ and $\boldsymbol{\kappa}''(\boldsymbol{b}) = (\boldsymbol{\kappa}_1''(\boldsymbol{b}_1)^T, \ldots, \boldsymbol{\kappa}_n''(\boldsymbol{b}_n)^T)^T$, and the functions $\boldsymbol{\kappa}_i'(\boldsymbol{b}_i)$ and $\boldsymbol{\kappa}_i''(\boldsymbol{b}_i)$, $i = 1, \ldots, n$, are given in (5.9).

Step2 : By one-step Newton-Rapshon iteration, the $(k+1)$-th estimate is calculated as $\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} - \big[S_P'(\widehat{\boldsymbol{\theta}}^{(k)T})\big]^{-1}\big[S_P(\widehat{\boldsymbol{\theta}}^{(k)T})\big]^T$, where $S_P(\boldsymbol{\theta})$ is the score equation for $\boldsymbol{\theta}$ from the penalized log-likelihood and $S_P'(\boldsymbol{\theta})$ is the first derivative of $S_P(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. With $(\widehat{\boldsymbol{\theta}}^{(k+1)}, \widetilde{\boldsymbol{b}}^{(k+1)})$, the $(k+1)$-th Breslow-type estimate of the baseline cumulative hazard for the s-th stratum is obtained as an empirical function which has jumps only at the observed failure time,

$$\begin{aligned}\Lambda_s^{(k+1)}(t) &= \Lambda_s^{(k+1)}(t; \widehat{\boldsymbol{\theta}}^{(k+1)}, \widetilde{\boldsymbol{b}}^{(k+1)}) \\ &= \sum_{i:V_i \le t} \frac{\Delta_i I(S_i = s)}{\sum_{l:V_l \ge V_i} \exp\big\{\widetilde{\boldsymbol{Z}}_l(V_i)(\widehat{\psi}^{(k+1)} \circ \widetilde{\boldsymbol{b}}_l^{(k+1)}) + \boldsymbol{Z}_l(V_i)\widehat{\gamma}^{(k+1)}\big\} I(S_l = s)}. \end{aligned} \quad (5.15)$$

For variance estimation of $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Lambda}}(t))$, we adopt the observed information matrix via Louis (1982) formula and conduct the Expectation step used in the maximum likelihood approach with the estimates by the penalized likelihood method. For the numerical calculation of the observed information matrix, we consider $\Lambda_s\{V_i\}$, the jump size of $\Lambda_s(t)$ at V_i belonging to stratum s for which $\Delta_i = 1$, instead of $\lambda_s(V_i)$. That is, $\boldsymbol{\Lambda}\{\cdot\} = (\boldsymbol{\Lambda}_1^T\{\cdot\}, \ldots, \boldsymbol{\Lambda}_S^T\{\cdot\})^T$ with $\boldsymbol{\Lambda}_s\{\cdot\} = (\Lambda\{T_{s1}\}, \ldots, \Lambda\{T_{sm_s}\})^T$ for m_s failure times among n_s subjects $(0 \le m_s \le n_s)$ of the s-th stratum, $s = 1, \ldots, S$. Then, by the Louis (1982) formula,

$$
\begin{aligned}
\mathrm{I}(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}) \;=\; & \mathrm{E}_{b|Y,V}[B_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})|\boldsymbol{Y}, \boldsymbol{V}] \\
& - \mathrm{E}_{b|Y,V}[U_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b}) U_c^T(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})|\boldsymbol{Y}, \boldsymbol{V}] \\
& + \mathrm{E}_{b|Y,V}[U_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})]\, \mathrm{E}_{b|Y,V}[U_c^T(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})],
\end{aligned}
$$

where $U_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ and $B_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$ are the first derivative vector and the negative of the second derivative matrix for the complete data log-likelihood $l_c(\boldsymbol{\theta}, \boldsymbol{\Lambda}\{\cdot\}; \boldsymbol{Y}, \boldsymbol{V}, \boldsymbol{b})$, respectively. For subject i with $S_i = s$, given observations and the penalized likelihood estimate $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s)$, we calculate the following conditional expectation of a known function $q(\boldsymbol{b}_i)$ needed in the observed information matrix,

$$
\begin{aligned}
E\big[q(\boldsymbol{b}_i)|\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s\big] \;&=\; \frac{\int_{\boldsymbol{b}_i} q(\boldsymbol{b}_i) f\big(\boldsymbol{Y}_i, V_i|\boldsymbol{b}_i, \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s\big) f\big(\boldsymbol{b}_i|\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s\big) d\boldsymbol{b}_i}{\int_{\boldsymbol{b}_i} f\big(\boldsymbol{Y}_i, V_i|\boldsymbol{b}_i, \widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s\big) f\big(\boldsymbol{b}_i|\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s\big) d\boldsymbol{b}_i} \\
&=\; \frac{\int_{\boldsymbol{z}_G} q\big(R(\boldsymbol{z}_G)\big) K(\boldsymbol{z}_G) \exp\{-\boldsymbol{z}_G^T \boldsymbol{z}_G\} d\boldsymbol{z}_G}{\int_{\boldsymbol{z}_G} K(\boldsymbol{z}_G) \exp\{-\boldsymbol{z}_G^T \boldsymbol{z}_G\} d\boldsymbol{z}_G},
\end{aligned} \tag{5.16}
$$

where $\boldsymbol{z}_G$ follows a multivariate Gaussian distribution with mean zero, $\boldsymbol{z}_G = R^{-1}(\boldsymbol{b}_i)$, $K(\boldsymbol{z}_G) = \exp\{\boldsymbol{z}_G^T \boldsymbol{z}_G\}\ f\big(\boldsymbol{Y}_i, V_i|R(\boldsymbol{z}_G), \boldsymbol{\theta}^{(k)}, \Lambda_s^{(k)}\big)\ f\big(R(\boldsymbol{z}_G)|\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_s\big)$, and Gauss-Hermite Quadrature numerical approximation is used for the calculation of integration. Note that, in (5.16), the functions of $R(\cdot)$ and $K(\cdot)$ have different expressions for different longitudinal distributions.

The proposed penalized likelihood approach for simultaneous modeling can be applied to all generalized linear mixed models of longitudinal outcomes. Next, we provide the expressions of the penalized log-likelihood and relevant equations for continuous and binary longitudinal outcomes with survival time.

Ex 1. Continuous longitudinal data with Normal distribution and survival time

Continuous longitudinal outcomes following a normal distribution has $A(D_i(t_j; \phi)) =$

σ_y^2, $B_{ij}(\boldsymbol{\beta};\boldsymbol{b}_i) = \left(\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\right)^2/2$, and $C(Y_{ij};D_i(t_j;\phi)) = -\left(y^2/\sigma_y^2+\log(\sigma_y^2)+\log(2\pi)\right)/2$ in (5.1), where σ_y^2 is the variance of longitudinal outcomes given $\boldsymbol{b}_i$. Then, the $\boldsymbol{\kappa}_i'(\boldsymbol{b}_i)$ and $\boldsymbol{\kappa}_i''(\boldsymbol{b}_i)$, $i = 1,\ldots,n$, used in Step 1 are

$$
\begin{aligned}
&\boldsymbol{\kappa}_i'(\boldsymbol{b}_i)\\
&= -\Bigg[\sum_{j=1}^{n_i}\frac{1}{\sigma_y^2}\left(Y_{ij}-\boldsymbol{X}_{ij}\boldsymbol{\beta}-\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\right)\widetilde{\boldsymbol{X}}_{ij}+\sum_{s=1}^{S}I(S_i=s)\Bigg(\Delta_i(\widetilde{\boldsymbol{Z}}_i(V_i)\circ\boldsymbol{\psi}^T)\\
&\qquad -\int_0^{V_i}\exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_i)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}(\widetilde{\boldsymbol{Z}}_i(u)\circ\boldsymbol{\psi}^T)d\Lambda_s(u)\Bigg)\Bigg]+\boldsymbol{b}_i^T\boldsymbol{\Sigma}_b^{-1}
\end{aligned}
$$

and

$$
\begin{aligned}
&\boldsymbol{\kappa}_i''(\boldsymbol{b}_i)\\
&= -\Bigg[\sum_{j=1}^{n_i}\left(-\frac{1}{\sigma_y^2}\right)\widetilde{\boldsymbol{X}}_{ij}^T\widetilde{\boldsymbol{X}}_{ij}\\
&\quad +\sum_{s=1}^{S}I(S_i=s)\left(-\int_0^{V_i}\exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_i)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}(\widetilde{\boldsymbol{Z}}_i^T(u)\circ\boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(u)\circ\boldsymbol{\psi}^T)d\Lambda_s(u)\right)\Bigg]\\
&\quad +\boldsymbol{\Sigma}_b^{-1}.
\end{aligned}
$$

In Step2, the penalized log-likelihood (5.14) has the following form for continuous longitudinal outcomes from a normal distribution and survival time,

$$
\begin{aligned}
&l_P(\boldsymbol{\theta},\boldsymbol{\Lambda})\\
&= \sum_{i=1}^{n}\Bigg[-\frac{1}{2}\log\left|\boldsymbol{I}_{d_b}-\boldsymbol{\Sigma}_b\left(\mathrm{E}\left[\widetilde{\boldsymbol{X}}_i^T\boldsymbol{W}_i\widetilde{\boldsymbol{X}}_i\right]-\sum_{s=1}^{S}I(S_i=s)\,\mathrm{E}\left[(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(V_i)\circ\boldsymbol{\psi}^T)\Delta_i\right]\right)\right|\\
&\qquad -\sum_{j=1}^{n_i}\frac{1}{2\sigma_y^2}\left(Y_{ij}-\boldsymbol{X}_{ij}\boldsymbol{\beta}-\widetilde{\boldsymbol{X}}_{ij}\widetilde{\boldsymbol{b}}_i\right)^2-\frac{n_i}{2}\log\left(2\pi\sigma_y^2\right)\\
&\qquad +\sum_{s=1}^{S}I(S_i=s)\Bigg[\Delta_i\log(\lambda_s(V_i))+\Delta_i[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi}\circ\widetilde{\boldsymbol{b}}_i)+\boldsymbol{Z}_i(V_i)\boldsymbol{\gamma}]\\
&\qquad\qquad -\int_0^{V_i}\exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\widetilde{\boldsymbol{b}}_i)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}d\Lambda_s(u)\Bigg]-\frac{1}{2}\widetilde{\boldsymbol{b}}_i^T\boldsymbol{\Sigma}_b^{-1}\widetilde{\boldsymbol{b}}_i\Bigg)\Bigg]. \qquad (5.17)
\end{aligned}
$$

For the expected values in (5.17), we use the original expressions of $\boldsymbol{W}$ and Δ_i, evaluated at the parameter estimates at the previous iteration and the random effects estimates from Step 1 at the current iteration, as $\widehat{\boldsymbol{W}}$ and $\widehat{\Delta}_i$, which are

$$-\sum_{j=1}^{n_i} \frac{1}{\widehat{\sigma}_y^2} \widetilde{\boldsymbol{X}}_{ij}^T \widetilde{\boldsymbol{X}}_{ij} \quad \text{and} \quad \exp\{\widetilde{\boldsymbol{Z}}_i(V_i)(\widehat{\boldsymbol{\psi}} \circ \widetilde{\boldsymbol{b}}_i) + \boldsymbol{Z}_i(V_i)\widehat{\boldsymbol{\gamma}}\}\widehat{\Lambda}_s(V_i),$$

respectively. On the other hand, the observed variance of longitudinal outcomes and the observed event for each subject also may be used as $\widehat{\sigma}_y^2$ and $\widehat{\Delta}_i$. $S_P(\boldsymbol{\theta})$ is obtained by differentiating (5.17) with respect to $\boldsymbol{\theta}$, and $S_P'(\boldsymbol{\theta})$ is the derivative of $S_P(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The Breslow-type estimator of the baseline cumulative hazard for the s-th stratum has the same expression given in (5.15) for all different longitudinal distributions.

Ex 2. Binary longitudinal data and survival time

Logistic distribution has $A(D_i(t_j;\phi)) = 1$, $B_{ij}(\boldsymbol{\beta};\boldsymbol{b}_i) = \log\left(1+\exp\left\{\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\right\}\right)$, and $C(Y_{ij}; D_i(t_j;\phi)) = 0$ in (5.1). Thus, the $\boldsymbol{\kappa}_i'(\boldsymbol{b}_i)$ and $\boldsymbol{\kappa}_i''(\boldsymbol{b}_i)$, $i = 1,\ldots,n$, in Step 1 are

$$\begin{aligned}
&\boldsymbol{\kappa}_i'(\boldsymbol{b}_i) \\
&= -\Bigg[\sum_{j=1}^{n_i}\left(Y_{ij} - \frac{\exp\{\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\}}{1+\exp\{\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\}}\right)\widetilde{\boldsymbol{X}}_{ij} + \sum_{s=1}^{S} I(S_i=s)\Bigg(\Delta_i(\widetilde{\boldsymbol{Z}}_i(V_i)\circ\boldsymbol{\psi}^T) \\
&\qquad -\int_0^{V_i}\exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_i)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}(\widetilde{\boldsymbol{Z}}_i(u)\circ\boldsymbol{\psi}^T)d\Lambda_s(u)\Bigg)\Bigg] + \boldsymbol{b}_i^T\boldsymbol{\Sigma}_b^{-1}
\end{aligned}$$

and

$$\begin{aligned}
&\boldsymbol{\kappa}_i''(\boldsymbol{b}_i) \\
&= -\Bigg[\sum_{j=1}^{n_i}\left(-\frac{\exp\{\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\}}{(1+\exp\{\boldsymbol{X}_{ij}\boldsymbol{\beta}+\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{b}_i\})^2}\right)\widetilde{\boldsymbol{X}}_{ij}^T\widetilde{\boldsymbol{X}}_{ij} \\
&\quad + \sum_{s=1}^{S} I(S_i=s)\left(-\int_0^{V_i}\exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\boldsymbol{b}_i)+\boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}(\widetilde{\boldsymbol{Z}}_i^T(u)\circ\boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(u)\circ\boldsymbol{\psi}^T)d\Lambda_s(u)\right)\Bigg] \\
&\quad +\boldsymbol{\Sigma}_b^{-1}.
\end{aligned}$$

In Step 2, the penalized log-likelihood (5.14) has the following form for binary longitudinal outcomes and survival time,

$$
\begin{aligned}
& l_P(\boldsymbol{\theta}, \boldsymbol{\Lambda}) \\
&= \sum_{i=1}^{n}\Bigg[-\frac{1}{2}\log\left|\boldsymbol{I}_{d_b} - \boldsymbol{\Sigma}_b\left(\mathrm{E}\left[\widetilde{\boldsymbol{X}}_i^T \boldsymbol{W}_i \widetilde{\boldsymbol{X}}_i\right] - \sum_{s=1}^{S} I(S_i = s)\,\mathrm{E}\left[(\widetilde{\boldsymbol{Z}}_i^T(V_i)\circ\boldsymbol{\psi})(\widetilde{\boldsymbol{Z}}_i(V_i)\circ\boldsymbol{\psi}^T)\Delta_i\right]\right)\right| \\
&\quad + \sum_{j=1}^{n_i}\left[Y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\widetilde{\boldsymbol{b}}_i) - \log\left(1 + \exp\left\{\boldsymbol{X}_{ij}\boldsymbol{\beta} + \widetilde{\boldsymbol{X}}_{ij}\widetilde{\boldsymbol{b}}_i\right\}\right)\right] \\
&\quad + \sum_{s=1}^{S} I(S_i = s)\Bigg[\Delta_i \log(\lambda_s(V_i)) + \Delta_i[\widetilde{\boldsymbol{Z}}_i(V_i)(\boldsymbol{\psi}\circ\widetilde{\boldsymbol{b}}_i) + \boldsymbol{Z}_i(V_i)\boldsymbol{\gamma}] \\
&\qquad - \int_0^{V_i} \exp\{\widetilde{\boldsymbol{Z}}_i(u)(\boldsymbol{\psi}\circ\widetilde{\boldsymbol{b}}_i) + \boldsymbol{Z}_i(u)\boldsymbol{\gamma}\}d\Lambda_s(u)\Bigg] - \frac{1}{2}\widetilde{\boldsymbol{b}}_i^T\boldsymbol{\Sigma}_b^{-1}\widetilde{\boldsymbol{b}}_i\Bigg)\Bigg]. \qquad (5.18)
\end{aligned}
$$

For the expected values in (5.18), we use the original expressions of $\widetilde{\boldsymbol{X}}_i^T \boldsymbol{W}_i \widetilde{\boldsymbol{X}}_i$ and Δ_i, evaluated at the parameter estimates at the previous iteration and the random effects estimates from Step 1 at the current iteration, as $\widetilde{\boldsymbol{X}}_i^T \widehat{\boldsymbol{W}} \widetilde{\boldsymbol{X}}_i$ and $\widehat{\Delta}_i$, which are

$$
-\sum_{j=1}^{n_i} \frac{\exp\{\boldsymbol{X}_{ij}\widehat{\boldsymbol{\beta}} + \widetilde{\boldsymbol{X}}_{ij}\widetilde{\boldsymbol{b}}_i\}}{(1 + \exp\{\boldsymbol{X}_{ij}\widehat{\boldsymbol{\beta}} + \widetilde{\boldsymbol{X}}_{ij}\widetilde{\boldsymbol{b}}_i\})^2}\widetilde{\boldsymbol{X}}_{ij}^T\widetilde{\boldsymbol{X}}_{ij} \quad \text{and} \quad \exp\{\widetilde{\boldsymbol{Z}}_i(V_i)(\widehat{\boldsymbol{\psi}}\circ\widetilde{\boldsymbol{b}}_i) + \boldsymbol{Z}_i(V_i)\widehat{\boldsymbol{\gamma}}\}\widehat{\Lambda}_s(V_i),
$$

respectively. On the other hand, since the original expression of $\widetilde{\boldsymbol{X}}_i^T \boldsymbol{W}_i \widetilde{\boldsymbol{X}}_i$ is same as $-n_i \,\mathrm{Var}(Y_{ij}|\boldsymbol{b}_i)$, the observed variance of longitudinal outcomes of each subject may be used as $\widehat{\mathrm{Var}}(Y_{ij}|\boldsymbol{b}_i)$. Likewise, the individual observed event also may be used as $\widehat{\Delta}_i$. $S_P(\boldsymbol{\theta})$ and $S_P'(\boldsymbol{\theta})$ are the first and second derivatives of (5.18) with respect to $\boldsymbol{\theta}$.

5.4 Simulation Studies

In this section, through simulation studies, we compare numerical performances on the computing time, bias, and mean squared error (MSE) of the penalized likelihood method and the EM algorithm used in maximum likelihood estimation for the simultaneous modeling of binary longitudinal outcomes and survival time with a random

intercept.

We assume that Y_{ij} is a binary outcome following

$$P(Y_{ij} = y_{ij}|b_i) = \exp\left\{y_{ij}\eta_{ij} - \log(1 + \exp\{\eta_{ij}\})\right\}, \qquad y_{ij} = 0,\ 1,$$

with $\eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + b_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3ij} + b_i$ for $j = 1, \ldots, n_i$, and

$$h(t|b_i) = \lambda(t)\exp\{\psi b_i + \boldsymbol{Z}_i(t)\boldsymbol{\gamma}\} = \lambda(t)\exp\{\psi b_i + \gamma_1 Z_{1i} + \gamma_2 Z_{2i}\},$$

where $b_i \sim N(0, \sigma_b^2)$, $X_{1i} \equiv Z_{1i}$ are simulated from a Bernoulli distribution with success probability being 0.5, and $X_{2i} \equiv Z_{2i}$ are simulated from the uniform distribution between 0 and 1. For the time at which longitudinal data are observed, we consider 4 different units of 0.3, 0.1, 0.05 and 0.03. The longitudinal data are generated for every unit of time, and thus X_{3ij}, the time at measurement, has the value of every unit ranging over 0 through 2.4. We consider $\psi = -0.1$ indicating negative dependency between longitudinal process and survival time model. The parameters in the two models are chosen as $\beta_0 = -1$, $\beta_1 = 1$, $\beta_2 = -0.5$, $\beta_3 = -0.2$, $\sigma_b^2 = 0.5$, $\psi = -0.1$, $\gamma_1 = -0.1$, $\gamma_2 = 0.1$, and $\lambda(t) = 1$. Censoring time is generated from the uniform distribution between 0.4 and 2.4, and the censoring proportion is around 25~35%. We consider different sample sizes (n) of 200 and 400 with 1000 replications and different average numbers of longitudinal observations per subject (n_i) which are 4, 8, 15 and 25. For the estimated baseline cumulative hazard function, we consider three fixed time points of 0.9, 1.4, and 1.9. Table 5.1 and Table 5.2 show the simulation results of maximum likelihood estimation (MLE) and maximum penalized likelihood estimation (MPLE) for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma_b^2, \psi, \boldsymbol{\gamma}^T)^T$ and baseline cumulative hazards at the given three time points in the simultaneous modeling of binary longitudinal outcomes and survival time with sample sizes of 200 and 400, respectively. In Table 5.1 and Table 5.2, "n_i" is the average number of lon-

Table 5.1: Summary of simulation results from maximum likelihood estimation (MLE) and maximum penalized likelihood estimation (MPLE) in the simultaneous modeling of binary longitudinal outcomes and survival time (n=200).

			MLE						MPLE					
n_i	Par.	True	Est.	Bias	SSD	ESE	MSE	CP	Est.	Bias	SSD	ESE	MSE	CP
4	β_0	-1.0	-1.013	- .013	.247	.243	.061	.948	- .932	.068	.226	.228	.056	.948
	β_1	1.0	1.004	.004	.203	.205	.041	.953	.928	- .072	.186	.190	.040	.933
	β_2	- .5	- .479	.021	.358	.349	.128	.944	- .444	.056	.330	.327	.112	.940
	β_3	- .2	- .201	- .001	.207	.209	.043	.960	- .191	.009	.193	.203	.037	.969
	γ_1	- .1	- .096	.004	.169	.174	.029	.964	- .098	.002	.170	.177	.029	.965
	γ_2	.1	.098	- .002	.301	.302	.091	.948	.100	.000	.301	.314	.091	.954
	ψ	- .1	- .111	- .011	.316	.316	.100	.980	- .131	- .031	.404	.472	.164	.989
	σ_b^2	.5	.516	.016	.204	.217	.042	.949	.360	- .140	.138	.172	.038	.997
	$\Lambda(.9)$	.9	.917	.017	.191	.184	.061	.945	.932	.032	.196	.200	.052	.952
	$\Lambda(1.4)$	1.4	1.446	.046	.302	.297	.043	.945	1.471	.071	.314	.328	.040	.959
	$\Lambda(1.9)$	1.9	1.977	.077	.440	.442	.134	.958	2.013	.113	.460	.494	.122	.966
8	β_0	-1.0	- .994	.006	.201	.198	.040	.946	- .927	.073	.187	.188	.040	.928
	β_1	1.0	.990	- .010	.165	.168	.027	.954	.927	- .073	.154	.158	.029	.933
	β_2	- .5	- .504	- .004	.296	.288	.087	.948	- .471	.029	.277	.273	.078	.945
	β_3	- .2	- .206	- .006	.156	.156	.024	.953	- .199	.001	.148	.152	.022	.956
	γ_1	- .1	- .102	- .002	.179	.173	.032	.939	- .103	- .003	.179	.173	.032	.939
	γ_2	.1	.114	.014	.288	.300	.083	.962	.116	.016	.288	.305	.083	.965
	ψ	- .1	- .112	- .012	.230	.232	.053	.976	- .114	- .014	.266	.264	.071	.974
	σ_b^2	.5	.502	.002	.138	.142	.019	.961	.402	- .098	.106	.115	.021	.998
	$\Lambda(.9)$	.9	.906	.006	.176	.181	.040	.959	.913	.013	.178	.188	.035	.964
	$\Lambda(1.4)$	1.4	1.421	.021	.282	.289	.028	.958	1.432	.032	.287	.299	.025	.961
	$\Lambda(1.9)$	1.9	1.949	.049	.413	.428	.090	.965	1.964	.064	.419	.441	.081	.968
15	β_0	-1.0	- .989	.011	.168	.166	.028	.946	- .938	.062	.159	.159	.029	.931
	β_1	1.0	.997	- .003	.145	.142	.021	.943	.948	- .052	.137	.136	.021	.933
	β_2	- .5	- .508	- .008	.248	.245	.062	.950	- .481	.019	.235	.235	.055	.952
	β_3	- .2	- .201	- .001	.109	.113	.012	.958	- .198	.002	.105	.112	.011	.961
	γ_1	- .1	- .083	.017	.167	.172	.028	.955	- .084	.016	.168	.172	.028	.954
	γ_2	.1	.114	.014	.301	.299	.091	.951	.118	.018	.301	.301	.091	.953
	ψ	- .1	- .098	.002	.185	.190	.034	.956	- .090	.010	.200	.200	.040	.950
	σ_b^2	.5	.487	- .013	.098	.100	.010	.966	.424	- .076	.082	.085	.012	.944
	$\Lambda(.9)$	.9	.900	.000	.182	.179	.028	.955	.902	.002	.183	.182	.025	.957
	$\Lambda(1.4)$	1.4	1.400	.000	.297	.283	.021	.948	1.403	.003	.297	.287	.019	.953
	$\Lambda(1.9)$	1.9	1.914	.014	.431	.416	.062	.949	1.919	.019	.434	.422	.055	.952
25	β_0	-1.0	- .992	.008	.149	.150	.022	.949	- .951	.049	.142	.145	.023	.938
	β_1	1.0	.997	- .003	.132	.130	.017	.947	.957	- .043	.125	.125	.017	.934
	β_2	- .5	- .501	- .001	.223	.225	.050	.950	- .481	.019	.209	.217	.044	.964
	β_3	- .2	- .203	- .003	.086	.090	.007	.954	- .200	.000	.084	.089	.007	.960
	γ_1	- .1	- .101	- .001	.177	.172	.031	.941	- .098	.002	.174	.172	.030	.941
	γ_2	.1	.100	.000	.310	.299	.096	.943	.112	.012	.305	.300	.093	.947
	ψ	- .1	- .091	.009	.177	.169	.031	.947	- .084	.016	.185	.173	.034	.938
	σ_b^2	.5	.490	- .010	.083	.084	.007	.956	.446	- .054	.073	.073	.008	.931
	$\Lambda(.9)$	.9	.913	.013	.188	.182	.022	.941	.910	.010	.186	.183	.020	.944
	$\Lambda(1.4)$	1.4	1.428	.028	.305	.288	.018	.937	1.421	.021	.305	.289	.016	.932
	$\Lambda(1.9)$	1.9	1.958	.058	.454	.426	.053	.947	1.946	.046	.450	.426	.046	.946

Table 5.2: Summary of simulation results from maximum likelihood estimation (MLE) and maximum penalized likelihood estimation (MPLE) in the simultaneous modeling of binary longitudinal outcomes and survival time (n=400).

			MLE						MPLE					
n_i	Par.	True	Est.	Bias	SSD	ESE	MSE	CP	Est.	Bias	SSD	ESE	MSE	CP
4	β_0	-1.0	-1.003	- .003	.172	.170	.030	.949	- .924	.076	.158	.159	.031	.932
	β_1	1.0	1.004	.004	.145	.143	.021	.946	.929	- .071	.132	.133	.023	.916
	β_2	- .5	-.503	-.003	.248	.244	.061	.941	-.465	.035	.229	.229	.053	.944
	β_3	- .2	-.196	.004	.147	.147	.022	.952	-.186	.014	.137	.142	.019	.959
	γ_1	- .1	-.100	.000	.124	.121	.015	.936	-.101	-.001	.124	.121	.015	.938
	γ_2	.1	.114	.014	.213	.210	.046	.950	.116	.016	.213	.213	.046	.957
	ψ	- .1	-.096	.004	.205	.203	.042	.963	-.112	-.012	.262	.270	.069	.978
	σ_b^2	.5	.496	- .004	.136	.150	.018	.961	.349	- .151	.095	.119	.032	1.000
	$\Lambda(.9)$	.9	.898	- .002	.129	.126	.030	.948	.905	.005	.131	.131	.025	.950
	$\Lambda(1.4)$	1.4	1.406	.006	.208	.201	.021	.949	1.418	.018	.213	.208	.018	.953
	$\Lambda(1.9)$	1.9	1.916	.016	.302	.295	.062	.944	1.934	.034	.311	.306	.053	.951
8	β_0	-1.0	- .998	.002	.147	.139	.022	.938	- .930	.070	.138	.132	.024	.903
	β_1	1.0	1.001	.001	.120	.118	.014	.946	.937	- .063	.112	.111	.017	.904
	β_2	- .5	-.509	-.009	.205	.203	.042	.951	-.476	.024	.192	.192	.038	.944
	β_3	- .2	-.195	.005	.112	.109	.013	.939	-.188	.012	.107	.107	.012	.942
	γ_1	- .1	-.098	.002	.125	.120	.016	.953	-.098	.002	.125	.120	.016	.953
	γ_2	.1	.106	.006	.207	.209	.043	.948	.107	.007	.206	.210	.043	.950
	ψ	- .1	-.104	-.004	.155	.156	.024	.967	-.103	-.003	.178	.176	.032	.965
	σ_b^2	.5	.498	- .002	.093	.099	.009	.963	.401	- .099	.072	.080	.015	.937
	$\Lambda(.9)$	.9	.902	.002	.127	.126	.022	.943	.905	.005	.127	.128	.019	.942
	$\Lambda(1.4)$	1.4	1.413	.013	.206	.200	.015	.946	1.417	.017	.206	.203	.013	.949
	$\Lambda(1.9)$	1.9	1.924	.024	.299	.292	.043	.946	1.930	.030	.300	.296	.038	.949
15	β_0	-1.0	- .996	.004	.117	.117	.014	.946	- .944	.056	.110	.112	.015	.923
	β_1	1.0	1.000	.000	.099	.101	.010	.954	.949	- .051	.095	.096	.012	.925
	β_2	- .5	-.504	-.004	.172	.173	.030	.955	-.477	.023	.163	.166	.027	.955
	β_3	- .2	-.199	.001	.081	.080	.006	.943	-.196	.004	.078	.079	.006	.947
	γ_1	- .1	-.100	.000	.118	.120	.014	.962	-.100	.000	.118	.120	.014	.964
	γ_2	.1	.108	.008	.208	.209	.043	.955	.110	.010	.208	.209	.043	.954
	ψ	- .1	-.099	.001	.128	.130	.016	.959	-.089	.011	.140	.136	.020	.953
	σ_b^2	.5	.495	- .005	.070	.071	.005	.959	.431	- .069	.058	.060	.008	.886
	$\Lambda(.9)$	.9	.899	- .001	.126	.126	.014	.955	.901	.001	.126	.127	.012	.955
	$\Lambda(1.4)$	1.4	1.405	.005	.198	.199	.010	.951	1.408	.008	.198	.200	.009	.953
	$\Lambda(1.9)$	1.9	1.917	.017	.289	.290	.030	.948	1.915	.015	.286	.292	.027	.951
25	β_0	-1.0	-1.004	- .004	.104	.106	.011	.949	- .961	.039	.099	.102	.011	.939
	β_1	1.0	.998	- .002	.093	.092	.009	.945	.954	- .046	.087	.088	.010	.921
	β_2	- .5	-.486	.014	.158	.159	.025	.943	-.467	.033	.150	.153	.024	.944
	β_3	- .2	-.198	.002	.063	.063	.004	.962	-.197	.003	.063	.063	.004	.964
	γ_1	- .1	-.101	-.001	.118	.120	.014	.961	-.100	.000	.118	.120	.014	.965
	γ_2	.1	.099	- .001	.217	.208	.047	.941	.103	.003	.213	.209	.046	.948
	ψ	- .1	-.096	.004	.115	.117	.013	.959	-.081	.019	.120	.120	.015	.957
	σ_b^2	.5	.493	- .007	.058	.059	.003	.964	.446	- .054	.051	.052	.006	.872
	$\Lambda(.9)$	.9	.911	.011	.135	.127	.011	.931	.910	.010	.132	.127	.010	.940
	$\Lambda(1.4)$	1.4	1.419	.019	.213	.200	.009	.933	1.414	.014	.207	.200	.008	.940
	$\Lambda(1.9)$	1.9	1.925	.025	.299	.291	.026	.947	1.925	.025	.294	.292	.023	.949

gitudinal outcomes per subject; "True" gives the true values of parameters; the middle 6 columns under "MLE" and the right 6 columns under "MPLE" are the results of the maximum likelihood estimates from the EM algorithm and the proposed maximum penalized likelihood estimates, repectively; the averages of the estimates are in "Est."; the averages of the bias estimates of the parameter estimates subtracted from true values are in "Bias"; the sample standard deviations from 1000 simulations are reported in "SSD"; "ESE" is the average of 1000 standard error estimates based on the observed information matrix; "MSE" gives the mean squared error calculated by adding the squared bias and the squared sample standard deviations; "CP" is the coverage proportion of 95% nominal confidence intervals based on the estimated standard error "ESE". Note that "ESE" under "MPLE" is based on the observed information matrix obtained by maximum likelihood approach using the maximum penalized likelihood estimates. Satterthwaite method is used for the coverage proportion of σ_b^2.

From Table 5.1 and Table 5.2, we can see that the bias of the proposed MPLE is small for most cases although it is bigger than the MLE's, but overall the bias of the MPLE decreases for the larger number of longitudinal observations per subject and the larger sample size like the MLE's does. On the other hand, the estimate of σ_b^2 of the MPLE is smaller than its true value showing the biggest bias, but it is improved soon being close to the true values as the number of longitudinal observations per subject increases. It is already known that the penalized quasi-likelihood (PQL) used for GLMMs tends to underestimate somewhat the variance components when applied to clustered binary data but the situation improves rapidly for binomial observations having denominators greater than one (Breslow and Clayton, 1993). The result from our simulation studies conforms this fact. For both MLE and MPLE, the estimated standard errors calculated from the observed information matrix are close to the sample standard deviations from the 1000 estimates. They decrease over the number of

longitudinal observations per subject except for the baseline cumulative hazards estimates, and they also decrease as sample size increases. The MPLE has smaller sample standard deviations and estimated standard errors than MLE for most cases. As for the mean squared error representing both bias and sample standard deviation together, the mean squared error of the MPLE appears to be smaller than or close to the MLE's. The mean squared errors from both MLE and MPLE decrease as the number of longitudinal observations per subject and sample size increase. The 95% confidence interval coverage rates are close to 0.95 except those for ψ of both MLE and MPLE with the small numbers of longitudinal observations per subject (n_i=4 and 8) of the small sample size (n=200), for σ_b^2 of the MPLE with the small number of longitudinal observations per subject (n_i=4 and 8) of the small sample size (n=200), and for σ_b^2 of the MPLE with the very small or large number of longitudinal observations per subject (n_i=4, 15 and 25) of the large sample size (n=400). For both MLE and MPLE, the coverage rate of the parameter ψ is recovered for the large number of longitudinal observations per subject and the large sample size. Thus, with small number of longitudinal observations per subject and small sample size, the test for ψ is conservative, which strengthens the test results when rejecting the null ($\psi = 0$), and the type I error becomes closer to the nominal level as the number of longitudinal observations per subject and sample size increase. While the high coverage rate of σ_b^2 of the MLE is improved for both large number of longitudinal observations per subject and large sample size, the coverage rate of σ_b^2 of the MPLE appears to be improved for the large number of longitudinal observations per subject with small sample size and the small number of longitudinal observations per subject with the large sample size. With the small sample size of 200 of Table 5.1, the high coverage rates of σ_b^2 of the MPLE at the small numbers of longitudinal observations per subject (n_i=4 and 8) are recovered at the large numbers of longitudinal observations per subject (n_i=15 and 25). On the other hand, with the

relatively small number of longitudinal observations per subject (n_i=8), the high coverage rate of σ_b^2 of the MPLE shown at the small sample size of 200 in Table 5.1 is improved for the large sample size of 400 in Table 5.1. In additional simulation studies conducted with the larger sample size of 800 whose results are not provided in this paper, the high coverage rates of σ_b^2 of the MPLE shown at the smallest number of longitudinal observations per subject (n_i=4) with sample sizes of 200 and 400 in both Tables 5.1 and 5.2 actually reached 95% nominal level for the sample size of 800. Figure 5.1 shows the ratios of mean squared errors (MSEs) of the proposed MPLE to the MLE with sample sizes of 200 and 400 for the parameters of predictors in longitudinal and hazard models. This figure confirms the results provided in Table 5.1 and Table 5.2 in that all plots indicate the ratios of mean squared errors are close to 1 which implies the proposed MPLE provides the mean squared errors close to the MLE's. Figure 5.2 shows the ratios of user times of the proposed MPLE to the MLE with sample sizes of 200 and 400. Both plots show the proposed MPLE is more efficient reducing about 70% of the computing time of the MLE over all different numbers of longitudinal outcomes per subject and sample sizes.

5.5 Analysis of the CHANCE Study

The Carolina Head and Neck Cancer Study (CHANCE) is a population based epidemiologic study conducted at 60 hospitals in 46 counties in North Carolina from 2002 through 2006 (Divaris *et al.* 2010). Patients were diagnosed with head and neck cancer (oral, pharynx, and larynx cancer) from 2002–2006. Their survival status was collected up to 2007 and QoL was evaluated over time for three years after diagnosis. QoL information was collected through questionnaires. Based on summary scores of the five domains of self-perceived quality of life including Physical Well-Being (PWB), Social/Family Well-Being (SWB), Emotional Well-Being (EWB), Functional Well-Being

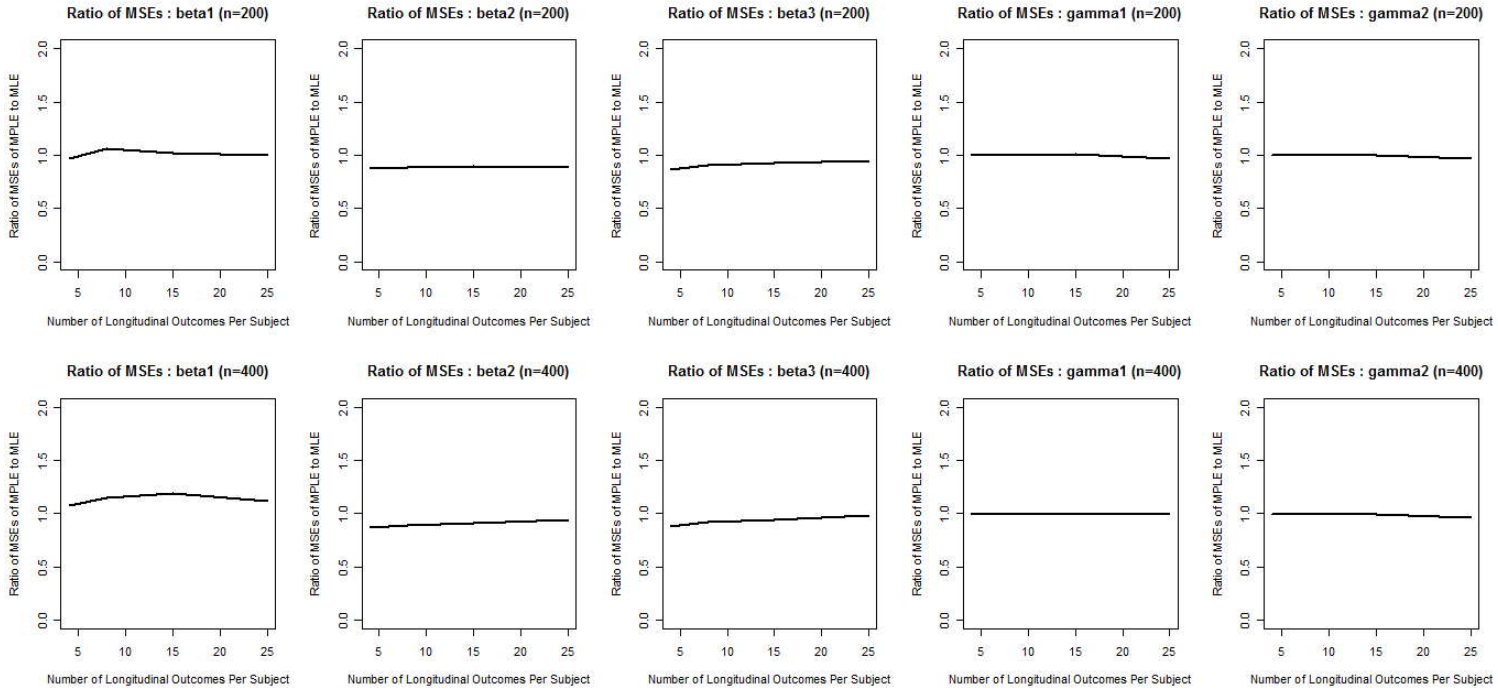

Figure 5.1: Plot of ratios of mean squared errors (MSEs) of maximum penalized likelihood estimator (MPLE) to maximum likelihood estimator (MLE) for parameters of predictors in longitudinal and hazard models (n=200, 400)

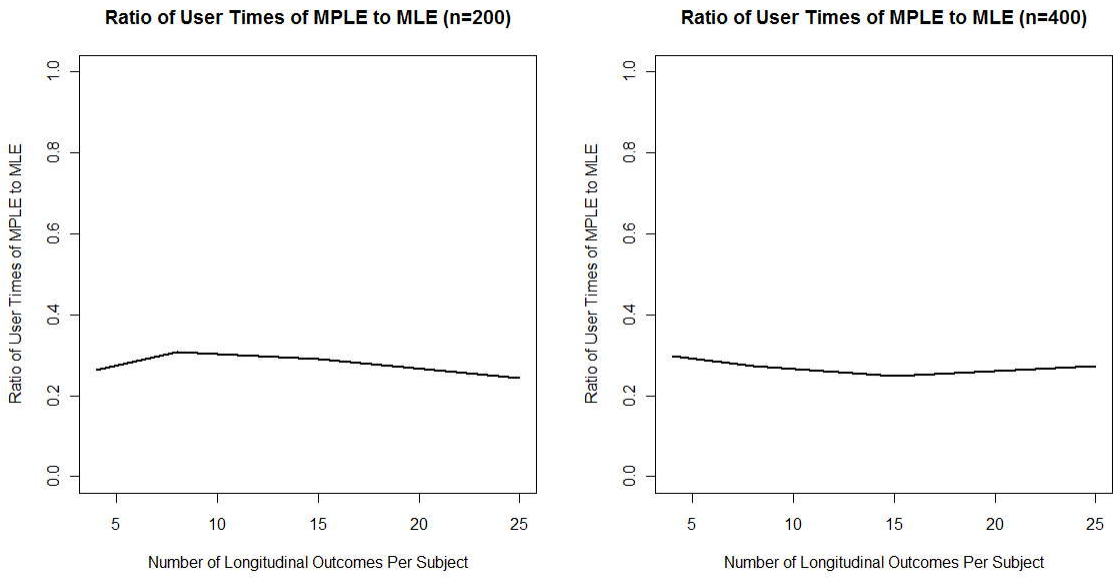

Figure 5.2: Plot of ratios of user times of maximum penalized likelihood estimator (MPLE) to maximum likelihood estimator (MLE) (n=200, 400)

(FWB) and Head and Neck Cancer Specific symptoms (HNCS), patient's QoL information was classified into satisfaction or dissatisfaction with life. Survival time is defined as the time to death from diagnosis. Demographic and life style characteristics, medical histories and clinical factors are also collected. Ending in December 2009 and excluding the patients with missing data, information on QoL has been obtained from 554 head and neck cancer patients in the analysis. Based on the death information through 2007 available from the National Death Index (NDI), 85 of 554 patients died and the censoring rate is 85%. The number of observations per patient ranges 1 to 3 with average of 1.93. It is of interest to elucidate the variables which are associated with both QoL satisfaction and survival time for patients with head and neck cancer. In particular, we are interested in the comparison between African-Americans and Whites since it is known that African-Americans have a higher incidence of head and neck cancer and worse survival than Whites. The longitudinal QoL satisfaction outcomes and survival time are correlated within a patient, and this dependency should be taken into account

in the analysis.

We apply both approaches of the maximum likelihood estimation (MLE) and maximum penalized likelihood estimation (MPLE) to Head and Neck Cancer Specific symptoms (HNCS) among QoL domains with survival time. Longitudinal HNCS QoL outcomes are binary measurements with 1 ("satisfied") and 0 ("dissatisfied"). In both longitudinal QoL and hazards models, we consider race (African-Americans, Whites), the number of 12 oz. beers consumed per week (None, <1, 1–4, 5–14, 15–29, ≥ 30), household income (0–10K, 20–30K, 40–50K, ≤ 60K), surgery (Yes/No), radiation therapy (Yes/No), chemotherapy (Yes/No), primary tumor site (Oral & Pharyngeal, Laryngeal) and tumor stage (I, II, III, IV) as categorical, and age at diagnosis (range: 24–80), the number of persons supported by household income (range: 1–5), body mass index (BMI) (range: 15.66–56.28) and the total number of medical conditions reported (range: 0–6) as continuous. Additionally, 2 interactions with race, i.e. race × the total number of medical conditions reported and race × tumor site, are included in both models since we are particularly interested in the difference of QoL and survival between African American and White. Time at survey measurement is also included as a covariate for longitudinal outcomes. A random intercept for the dependence between the QoL satisfaction and the risk of death is included in both models, and assumed to follow a mixture of normal distributions. In Table 5.3, we compare the estimates and the estimated standard errors of the maximum likelihood estimation (MLE) and maximum penalized likelihood estimation (MPLE). From both "Est." and "ESE" columns, we see that MLE and MPLE provide similar estimates and estimated standard errors each other for the parameters of interest in longitudinal QoL and hazards models. On the other hand, the parameters of σ_b^2 and ψ, which denote the variance of random effects and the coefficient of random effects characterizing the dependence between longitudinal QoL and survival processes, respectively, have different estimates and estimated

Table 5.3: Analyses results from maximum likelihood estimation (MLE) and maximum penalized likelihood estimation (MPLE) for the Quality of Life and survival time for the CHANCE study

Parameter		Est.		ESE	
		MLE	MPLE	MLE	MPLE
HNCS QoL longitudinal model					
Intercept	β_0	.197	.206	.824	.922
Race (ref= White): African American	β_1	.562	.564	.345	.391
# of 12 oz. beers consumed per week (ref=30 or more)					
– None	β_2	.615	.618	.275	.315
– less than 1	β_3	.751	.706	.366	.409
– 1 to 4	β_4	1.259	1.253	.300	.337
– 5 to 14	β_5	1.062	1.072	.257	.294
– 15 to 29	β_6	.581	.577	.294	.336
Household income (ref= level1: 0–10K)					
– level2: 20–30K	β_7	- .268	- .254	.236	.271
– level3: 40–50K	β_8	.291	.309	.257	.293
– level4: ≥ 60K	β_9	1.181	1.162	.279	.313
Surgery (ref= No): Yes	β_{10}	- .029	- .032	.205	.234
Radiation therapy (ref= No): Yes	β_{11}	-1.179	-1.098	.299	.323
Chemotherapy (ref= No): Yes	β_{12}	.219	.197	.245	.280
Tumor site (ref=Oral & Pharyngeal): Laryngeal	β_{13}	- .225	- .218	.225	.255
Tumor stage (ref= I)					
– II	β_{14}	- .479	- .470	.306	.334
– III	β_{15}	-1.494	-1.418	.320	.358
– IV	β_{16}	-1.383	-1.342	.308	.343
Age at diagnosis	β_{17}	.012	.012	.009	.011
# of persons supported by household income	β_{18}	- .174	- .176	.088	.100
BMI	β_{19}	.042	.040	.015	.017
Total # of medical conditions reported	β_{20}	- .208	- .203	.092	.104
Race (African-American) × Tumor site (Laryngeal)	β_{21}	- .156	- .178	.438	.496
Race (African-American) × Total # of medical conditions reported	β_{22}	.088	.095	.197	.224
Time at survey measurement (years)	β_{23}	.243	.216	.067	.070
variance of random effects	σ_b^2	.317	1.037	.185	.394
Hazards model					
Random effect coefficient	ψ	-1.560	- .623	1.060	.285
Race (ref= White): African American	γ_1	.482	.411	.461	.384
# of 12 oz. beers consumed per week (ref=30 or more)					
– None	γ_2	- .866	- .795	.417	.354
– less than 1	γ_3	- .241	- .198	.444	.395
– 1 to 4	γ_4	- .915	- .845	.447	.389
– 5 to 14	γ_5	-1.180	-1.106	.409	.350
– 15 to 29	γ_6	- .616	- .568	.422	.372
Household income (ref= level1: 0–10K)					
– level2: 20–30K	γ_7	- .168	- .195	.321	.283
– level3: 40–50K	γ_8	- .898	- .852	.400	.353
– level4: ≥ 60K	γ_9	-1.359	-1.319	.446	.396
Surgery (ref= No): Yes	γ_{10}	- .489	- .501	.324	.282
Radiation therapy (ref= No): Yes	γ_{11}	- .454	- .468	.411	.361
Chemotherapy (ref= No): Yes	γ_{12}	.065	.052	.372	.334
Tumor site (ref=Oral & Pharyngeal): Laryngeal	γ_{13}	- .012	- .018	.333	.295
Tumor stage (ref= I)					
– II	γ_{14}	- .163	- .240	.489	.428
– III	γ_{15}	.302	.179	.506	.426
– IV	γ_{16}	1.239	1.086	.478	.375
Age at diagnosis	γ_{17}	.023	.017	.019	.008
# of persons supported by household income	γ_{18}	.086	.059	.137	.110
BMI	γ_{19}	.015	.011	.022	.016
Total # of medical conditions reported	γ_{20}	.277	.249	.134	.112
Race (African-American) × Tumor site (Laryngeal)	γ_{21}	.342	.339	.573	.505
Race (African-American) × Total # of medical conditions reported	γ_{22}	- .256	- .242	.270	.245

standard errors between the MLE and MPLE. This discrepancy of the MPLE from the MLE may be a numerical issue due to the small cluster size with the average of 1.93. In addition, the MPLE provides slightly bigger estimated standard errors than the MLE for the parameters in the longitudinal model while it appears in the reverse direction in hazards model. This also may be a numerical issue due to the small number of longitudinal outcomes per subject since the estimation in the longitudinal model is directly affected by the individual cluster size while the estimation in hazards model is not. Comparing the computing time spent on producing the results in Table 5.3, the proposed MPLE took only a sixth of the time the MLE did (62.83 and 361.78 seconds for MPLE and MLE respectively). This analysis result indicates that, even for the small cluster size, the proposed MPLE provides the similar results to those of the MLE for the parameters of interest taking less computing time than the MLE. In the studies with larger number of longitudinal outcomes per subject, the results of the MPLE are expected to be close to those of the MLE for all parameters with much better efficiency on calculation.

5.6 Concluding Remarks

In this paper, we have developed a more computationally efficient estimation procedure adopting a penalized likelihood based on Laplace approximation for the simultaneous modeling of longitudinal outcomes and survival time. Our proposed penalized likelihood estimation method is an effort to reduce the intensity on computation still providing the similar estimates to those by the EM algorithm of the maximum likelihood approach. Simulation studies indicated that the penalized likelihood approach performs as well as the EM algorithm of maximum likelihood approach, but only requires a fraction of the computing time. We also illustrated this comparison with the CHANCE data.

For the purpose of comparison, we also conducted the simulation studies of two ad-

ditional approximation methods which are using the Laplace approximation to the full log-likelihood in (5.10) and using the observed variance of the longitudinal outcomes and the observed event for the expected values in the penalized log-likelihood (5.18) for binary longitudinal outcomes and survival time. The simulation results indicated that the former provides the estimates more close to the maximum likelihood estimates but takes longer time than the two penalized likelihood procedures although its computing time is also less than the maximum likelihood approach. It is because, from the Laplace approximated full log-likelihood, all parameters in $|\boldsymbol{I}_{d_b} - \boldsymbol{\Sigma}_b \widetilde{l}''_{i|b}(\boldsymbol{\theta}, \Lambda_s)|$ are used for estimation and thus the calculation is more complicated than the penalized likelihood estimation methods. In the mean time, the latter provides the most biased estimates but takes the least computing time since it uses the most approximated expression of log-likelihood. Therefore, in terms of bias and computing time, the proposed penalized likelihood method using the original expressions for the expected values in the penalized likelihood appears to behave best among the three approximation methods.

In the simultaneous modeling considered in this paper, we assumed random effects to follow a Gaussian distribution with mean zero. However, it is unclear whether the normality assumption is truly satisfied in practice. Future work can include developing an approach to diminish computational intensity efficiently through the penalized likelihood approach for relaxing the normality assumption of random effects in the simultaneous modeling.

Chapter 6

SUMMARY AND FUTURE RESEARCH

In this dissertation, we have studied statistical methods for joint analysis of survival time and longitudinal data and particularly proposed the simultaneous modeling of the two different types of outcomes. Random effects are introduced to account for the dependence between longitudinal outcomes and survival time due to unobserved factors. Specifically, in terms of the distributional assumption of random effects, the following two scenarios were studied: 1) assuming a Gaussian process for random effects, 2) assuming the underlying distribution of random effects to be unknown.

This dissertation research is motivated by biomedical and public health applications where it is common that both longitudinal outcomes over time and survival endpoint are collected for the same subject along with the subject's characteristics or risk factors. Investigators are often interested in finding important variables for predicting both longitudinal outcomes and survival time which are correlated within a subject. This naturally led us to consider simultaneous models. Thus, the main contribution of this dissertation is to provide statistical methods which address the association of covariates to both longitudinal outcomes and survival time of interest while the dependence between the two outcomes is taken into account. Generalized linear mixed model was considered for longitudinal process in order to incorporate both categorical

and continuous longitudinal outcomes. A stratified proportional hazards model was assumed for survival time. The cumulative baseline hazard functions were also studied and Breslow-type estimates were proposed.

In Chapter 3, we have assumed random effects to follow a multivariate Gaussian process in the joint analysis and considered a nonparametric maximum likelihood estimation approach. The EM algorithm was adopted for estimation, and the proposed estimates performed well in finite samples under the various simulation settings considered. The variance estimates based on the observed information matrix approximated the true variance well in finite samples. In Chapter4, we have relaxed normality assumption for random effects in the joint analysis. Assuming the underlying distribution of random effects to be unknown, we used a mixture of Gaussian distributions as an approximation for the random effect distribution. We also developed a nonparametric maximum likelihood estimation method, and weights of the mixture components were estimated with model parameters using the EM algorithm. The proposed estimators were shown to have nice finite sample properties via simulation studies. The simulation studies conducted for robustness of the assumed mixture distribution indicated that, when the true distribution of random effects is not normal, the mixture of normal distributions well-approximate the random effect distribution by yielding the less biased estimates for the parameters of interest in longitudinal and hazards models and the more similar shaped density plot to the true distribution than no mixture. The number of mixture distributions was shown to be properly selected by AIC and BIC through simulation studies. For both maximum likelihood approaches with and without normality assumption of random effects studied in Chapters 3 and 4 respectively, the proposed estimators were proved to have desirable asymptotic properties such as consistency and asymptotic normality, and most of the proofs relied on the empirical processes theory. In Chapter 5, we have considered the penalized likelihood for

a more computationally efficient estimation procedure than the EM algorithm of the maximum likelihood approach with a Gaussian process for random effects. Simulation studies showed the proposed penalized likelihood approach reduced the computational intensity spending only a fraction of the computing time the EM algorithm took, still providing the similar estimates and mean squared errors to those by the EM algorithm.

All proposed methods in this dissertation were illustrated with the real-world data sets from the CHANCE study. We compared the results from the proposed simultaneous analysis and separate analyses in Chapters 3 and 4. Under both situations of normally distributed and distributional free random effects, the simultaneous analysis additionally identified more predictors for longitudinal quality of life and survival time than separate analyses. In Chapter 5, we compared the analysis results from the maximum likelihood approach and maximum penalized likelihood approach for assuming normality of random effects. The latter showed the remarkable reduction from the former's computing time while both produced the similar results in estimation.

The proposed methods in this dissertation research can be extended in several directions:

First, in this dissertation, we considered the generalized linear mixed model for longitudinal process to incorporate both categorical and continuous longitudinal outcomes. In the the joint analysis framework, relatively little work was done for categorical longitudinal data while continuous longitudinal data were studied by many authors. Our proposed approaches generalize previous work to general longitudinal outcomes and this work fills in some gaps in the joint modeling research. Then, future work can include considering generalization to mixed types of longitudinal outcomes.

Second, in some applications where sample size and the number of observations per subject are too large, the EM algorithm of maximum likelihood approach may be intensive on computation due to the integration of complete data likelihood over

random effects. Thus, in my dissertation, we considered a penalized likelihood for the simultaneous modeling with a Gaussian process of random effects. It will be also of interest to develop an approach to relieve computational burden efficiently through the penalized likelihood approach for relaxing the normality assumption of random effects in the simultaneous modeling.

Third, we considered one survival event in the simultaneous analysis proposed in this dissertation, but one may be interested in multivariate endpoints such as recurrent events, multiple disease outcome data and competing risk factors. Therefore, exploring the possibility of the extension of the proposed approaches to multivariate survival data would be worth pursuing.

Last, but not least, in real applications, the proportional hazards assumption considered in this dissertation may not always be true or one may be interested in modeling association from different aspects. A natural extension would be to consider other types of models for survival time including, but not limited to the proportional odds model, the accelerated failure time model, or the additive hazards model.

REFERENCES

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004), Examples in Which Misspecification of a Random Effects Distribution Reduces Efficiency, and Possible Remedies, *Computational Statistics and Data Analysis*, **47,** 639–653.

Albert, P.S. and Follmann, D.A. (2000), Modeling Repeated Count Data Subject to Informative Dropout, *Biometrics*, **56**, 667–677.

Albert, P.S. and Follmann, D.A. (2007), Random Effects and Latent Processes Approaches for Analyzing Binary Longitudinal Data with Missingness: a Comparison of Approaches Using Opiate Clinical Trial Data, *Statistical Methods in Medical Research*, **16**, 417–439.

Albert, P.S., Follmann, D.A., Wang, S.A., and Suh, E.B. (2002), A Latent Autoregressive Model for Longitudinal Binary Data subject to Informative Missingness, *Biometrics*, **58,** 631–642.

Andersen, P.K. and Gill, R.D. (1982), Cox's Regression Model for Counting Processes: A Large Sample Study, *The Annals of Statistics*, **10,** 1100–1120.

Andersen, P.K, Borgan, O., Gill, R.D. and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer–Verlag.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989), *Asymptotic Techniques for Use in Statistics*, London: Chapman and Hall.

Bartholomew, D.J. (1987), *Latent Variable Models and Factor Analysis*, Oxford University Press, New York.

Bartlett, R.F. and Sutradhar, B.C. (1999), On estimating equations for parameters in generalized linear mixed models with application to binary data, *Environmetrics*, **10,** 769–784.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press,

Baltimore.

Böhning, D. (1999), *Computer-Assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping and Others*, Number 81 in Monographs on Statistics and Applied Probability, Chapman & Hall/CRC.

Breslow, N.E. and Clayton, D.G., (1993), Approximate Inference in Generalized Linear Mixed Models, *Journal of American Statistical Association*, **88,** 125-134.

Breslow, N.E. and Day, N.E. (1980), *Statistical Methods in Cancer Research, Volume I.* IARC Scientific Publications No. 32. Lyon.

Breslow, N.E. and Lin, X. (1995), Bias Correction in Generalized Linear Mixed Models with a Single Component of Dispersion, *Biometrika*, **82,** 81–91.

Brown, E.R. and Ibrahim, J.G. (2003), A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data, *Biometrics*, **59,** 221–228.

Buzas, J.S. (1998), Unbaised Scores in Proportional Hazards Regression with Covariate Measurement Error, *Journal of Statistical Planning and Inference*, **67,** 247–257.

Cai, J. and Prentice, R.L (1995), Estimating Equations for Hazard Ratio Parameters Based on Correlated Failure Time Data, *Biometrika*, **82,** 151–164.

Cai, J. and Prentice, R.L. (1997), Regression Analysis for Correlated Failure Time Data, *Lifetime Data Analysis*, **3,** 197–213.

Caffo, B., Ming-Wen, A., and Rohde, C. (2007), Flexible Random Intercept Models for Binary Outcomes Using Mixtures of Normals, *Computational Statistics and Data Analysis*, **51,** 5220–5235.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman and Hall.

Cella, D.F. (1994), *Manual: Functional Assessment of Cancer Therapy (FACT) Scales and Functional Assessment of HIV Infection (FAHI) Scale*, Chicago: Rush-Presbyterian-St. Luke's Medical Center

Cella, D.F., Tulsky, D.S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., Silberman, M., Yellen, S.B., Winicour, P., and Brannon, J. (1993), The Functional Assessment

of Cancer Therapy Scale: Development and Validation of the General Measure, *Journal of Clinical Oncology*, **11,** 570–579

Clayton, D. and Cuzick, J. (1985), Multivariate Generalizations of the Proportional Hazards Model (with discussion), *Journal Royal Statistical Society A*, **148,** 82–117.

Clegg, L.X, Cai, J. and Sen, P.K. (1999), A Marginal Mixed Baseline Hazards Model for Multivariate Failure Time Data, *Biometrics*, **55,** 805–812.

Cox, D.R. (1972), Regression Models and Life–tables (with discussion), *Journal Royal Statistical Society B*, **34,** 187–220.

Cox, D.R. (1975), Partial Likelihood, *Biometrika*, **62,** 269–276.

Cox, D.R. and Oakes, D. (1984), *Analysis of Survival Data*, London: Chapman & Hall.

Crouch, A.C. and Spiegelman, E. (1990), The Evaluation of Integrals of the Form $\int f(t)\exp(-t^2)$ṭ: Application to Logistic–normal Models, *Journal of American Statistical Association*, **85,** 464–469.

Dang, Q.Y., Mazumdar, S., and Houck, P.R. (2008), Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes, *Computer Methods and Programs in Biomedicine*, **91**, 122–127.

D'Antionio, L.L., Zimmerman, G.J., Cella, D.F., and Long, S.A. (1996), Quality of Life and Functional Status Measures in Patients with Head and Neck Cancer, *Archives of Otolaryngology Head & Neck Surgery*, **122,** 482–487

Davidian, M. and Gallant, A.R. (1992), The Nonlinear Mixed Effects Model with a Smooth Random Effects Density. *Department of Statistics Technical Report, North Carolina State University* , Campus Box 8203, Raleigh, North Carolina 27695.

DeGruttola, V. and Tu, X.M. (1994), Modeling Progression of CD-4 Lymphocyte Count and Its Relationship to Survival Time, *Biometrics*, **50,** 1003–1014.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), Maximum Likelihood from Incomplete Data via the EM algorithm, *Journal Royal Statistical Society B*, **39,** 1–38.

Diggle, P.J. (1988), An Approach to the Analaysis of Repeated Measurements, *Bio-*

metrics, **44,** 959–971.

Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002), *Analysis of Longitudinal Data*, Oxford University Press 2nd Ed.

Ding, J. and Wang, J.L. (2008), Modeling Longitudinal Data with Nonparametric Multiplicative Random Effects Jointly with Survival Data, *Biometrics*, **64,** 546–556.

Divaris, K., Olshan, A.F., Smith, J., Bell, M.E., Weissler, M.C., Funkhouser, W.K., and Bradshaw, P.T. (2010), "Oral Health and Risk for Head and Neck Squamous Cell Carcinoma: the Carolina Head and Neck Cancer Study, *Cancer Causes Control*, **21,** 567–575.

Drum, M.L. and McCullagh, P. (1993), REML Estimation with Exact Covariance in the Logistic Mixed Model, *Biometrics*, **49,** 677–689.

Elashoff, R.M., Li, G., and Li, N. (2007), An Approach to Joint Analysis of Longitudinal Measurements and Competing Risks Failure Time Data, *Statistics in Medicine*, **26,** 2813–2835.

———(2008), A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types, *Biometrics*, **64,** 762–771.

Fang, F.M., Chien, C.Y., Kuo, S.C., Chiu, H.C., and Wang, C.J. (2004), Changes in quality of life of head-and-neck cancer patients following postoperative radiotherapy, *Acta Oncologica*, **43,** 571–578.

Faucett, C.L., Schenker, N., and Elashoff, R.M. (1998), Analysis of Censored Survival Data with Intermittently Observed Time-Dependent Binary Covariates, *Journal of American Statistical Association*, **93,** 427–437.

Faucett, C.J. and Thomas, D.C. (1996), Simultaneously Modeling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach, *Statistics in Medicine*, **15,** 1663–1685.

Fieuws, S., Spiessens, B., and Draney, K. (2004), *Mixture Models. In: De Boeck, P., Wilson, M. (Eds.), Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, Springer-Verlag, New York, Ch. 11, 317–340.

Firth, D. (1991), Generalized Linear Models, *Statistical Theory and Modelling*, eds. Hinkley, D.V. and Snell, E.J, London: Chapman and Hall, 55–82.

Flemming, T.R. and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, New York: Wiley.

Gallant, A.R. and Nychka, D.W. (1987), Semi-nonparametric Maximum Likelihood Estimation, *Econometrica*, **55,** 363–390.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004), Smooth Random Effects Distribution in a Linear Mixed Model, *Biometrics*, **60,** 945–953.

Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985), The Analysis of Binomial Data by a Generalized Linear Mixed Model, *Biometrika*, **72,** 593–599.

Green, P.J. (1987), Penalized Likelihood for General Semi-parametric Regression Models, *International Statistical Review*, **55,** 245–259.

Heagerty, P.J. and Kurland, B.F. (2001), Misspecified Maximum Likelihood Estimates and Generalised Linear Mixed Models, *Biometrika*, **88,** 973–985.

Henderson, R., Diggle, P., and Dobson, A. (2000), Joint Modeling of Longitudinal Measurements and Event Time Data, *Biometrics*, **4,** 465–480.

Hogan, J. and Laird, N. (1997), Mixture Models for the Joint Distribution of Repeated Measures and Event Times, *Statistics in Medicine*, **16,** 239–257.

Holloway, R.L., Hellewell, J.L., Marbella, A.M., Layde, P.M., Myers, K.B., and Campbell, B.H. (2005), Psychosocial effects in long-term head and neck cancer survivors, *Head and Neck – Journal for the Sciences and Specialties of the Head and Neck*, **27,** 281–288.

Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, New York: Springer.

Hsieh, F., Tseng, Y.K. and Wang, J.L. (2006), Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited, *Biometrics*, **62,** 1037–1043.

Hu, W., Li, G., and Li, N. (2009), A Bayesian Approach to Joint Analysis of Longitudinal Measurements and Competing Risks Failure Time Data, *Statistics in Medicine*, **28,** 1601–1619.

Huang, W., Zeger, S., Anthony J., and Garrett E. (2001), Latent Variable Model for Joint Analysis of Multiple Repeated Measures and Bivariate Event times, *Journal of American Statistical Association*, **96,** 906–914.

Huber, P., Ronchetti, E., and Victoria-Feser, M.P. (2004), "Estimation of generalized linear latent variable models", *Journal Royal Statistical Society B*, **66,** 893–908.

Jang, W. and Lim, J. (2009), A Numerical Study of PQL Estimation Biases in Generalized Linear Mixed Models Under Heterogeneity of Random Effects, *Communications in Statistics-Simulation and Computation*, **38,** 692–702.

Kalbfleish, J.D. and Prentice, R.L. (2002), *The Statistical Analysis of Failure Time Data*, Wiley, John & Sons 2nd Ed.

Klein, J.P. (1992), Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm, *Biometrics*, 48, 795–806.

Kleinman, K.P. and Ibrahim, J.G. (1998), A Semiparametric Bayesian Approach to the Random Effects Model, *Biometrics*, **54,** 921–938.

Komárek, A. and Lesaffre E. (2008a), "Generalized Linear Mixed Model with a Penalized Gaussian Mixture as a Random Effects Distribution, *Computational Statistics and Data Analysis*, 52, 3441–3458.

———(2008b), Bayesian Accelerated Failure Time Model with Multivariate Doubly Interval-Censored Data and Flexible Distributional Assumptions, *Journal of American Statistical Association*, **103,** 523–533.

———(2009), The Regression Analysis of Correlated Interval-censored Data: Illustration Using Accelerated Failure Time Models with Flexible Distributional Assumptions, *Statistical Modelling*, **9,** 299–319.

Laird N.M. (1978), Empirical Bayes Methods for Two-Way Contingency Tables, *Biometrika*, **65,** 581–590.

Laird, N. (1978), Nonparametric Maximum Likelihood Estimation of a Mixing Distribution, *Journal of American Statistical Association*, **73,** 805–811.

Laird, N.M. and Ware, J.H. (1982), Random Effets Models for Longitudinal Data, *Journal of American Statistical Association*, **73,** 963–974.

Lange, N. and Ryan, L. (1989), Assessing Normality in Random Effects Models, *The Annals of Statistics*, **17,** 624–642

Larsen K. (2004), Joint Analysis of Time-to-Event and Multiple Binary Indicators of Latent Classes, *Biometrics*, **60,** 85–92.

Lee, E.W., Wei, L.J., and Amato, D.A. (1992), Cox–Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations, *Survival Analysis: State of the Art. J. P. Klein and P.K. Goel (eds.)*, Kluwer Academic Publishers, 237–247.

Li, Y. and Lin, X. (2000), Covariate Measurement Errors in Frailty Models for Clustered Survival Data, *Biometrika*, **87,** 846–866.

Li, E., Wang, N. and Wang, N.Y. (2007), Joint Models for a Primary Endpoint and Multiple Longitudinal Covariate Processes, *Biometrics*, **63,** 1068–1078.

Liang, K.Y. and Zeger, S.L. (1986), Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, **73,** 13–22.

Liang, K.Y., Self, S.G., and Chang, Y.(1993), Modeling Marginal Hazards in Multivariate Failure Time Data, *Journal Royal Statistical Society B*, **55,** 441–453.

Lin, D.Y. (1994), Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach, *Statistics in Medicine*, **13,** 2233–2247.

Lin, D.Y. (2000), On Fitting Cox's Proportional Hazards Models to Survey Data, *Biometrika*, **87,** 37–47.

Lin, X. and Breslow, N.E. (1996), Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion, *Journal of American Statistical Association*, **91,** 1007–1016.

List, L.L., D'Antionio, L.L., Cella, D.F., Siston, A., Mumby, P., Haraf, D., and Vokes, E. (1996), The Performance Status Scale for Head and Neck Cancer Patients and the Functional Assessment of Cancer Therapy–Head and Neck (FACT–H&N) Scale: A Study of Utility and Validity, *Cancer*, **77,** 2294–2301

Localio, A.R., Berlin, J.A., and Ten Have T.R. (2006), Longitudinal and repeated cross-sectional cluster-randomization designs using mixed effects regression for binary

outcomes: Bias and coverage of frequentist and Bayesian methods, *Statistics in Medicine*, **25,** 2720–2736.

Louis, T.A. (1982), Finding the Observed Information Matrix when Using the EM Algorithm, *Journal Royal Statistical Society B*, **44,** 226–233

Masaoud, E. and Stryhn, H. (2010), A simulation study to assess statistical methods for binary repeated measures data, *Preventive Veterinary Medicine*, **93,** 81–97.

McDullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Chapman & Hall/ CRC 2nd Ed.

Nakamura, T. (1992), Proportional Hazards Models with Covariates Subject to Measurement Error, *Biometrika*, **48,** 829–838.

Nelson, K.P. and Leroux, B.G. (2008), Properties and comparison of estimation methods in a log-linear generalized linear mixed model, *Journal of Statistical Computation and Simulation*, **78,** 367–384.

Neuhaus, J.M., Hauck, W.W., and Kalbfleisch, J.D. (1992), The Effects of Mixture Distribution Misspecification When Fitting Mixed-Effects Logistic Models, *Biometrika*, **79,** 755–762.

Neyman, J. and Scott, E.L. (1948), Consistent estimates based on partially consistent observations, *Econometrica*, **16,** 1–32.

Nibu, K.I., Ebihara, Y., Ebihara, M., Kawabata, K., Onitsuka, T., Fujii, T., and Saikawa, M. (2010), Quality of life after neck dissection: a multicenter longitudinal study by the Japanese Clinical Study Group on Standardization of Treatment for Lymph Node Metastasis of Head and Neck Cancer, *International Journal of Clinical Oncology*, **15,** 33–38.

Oakes, D. (1989), Bivariate Survival Models Induced by Frailties, *Journal of American Statistical Association*, **84,** 487–493.

Parner, E. (1998), Asymptotic Theory for the Correlated Gamma-frailty Model, *The Annals of Statistics*, **26,** 183–214.

Patterson, H.D. and Thompson, R. (1974), Recovery of Interblock Information When Block Sizes are Unequal, *Biometrika*, **58,** 545–554.

Pawitan, Y. and Self, S. (1993), Modeling Disease Marker Processes in AIDS, *Journal of American Statistical Association*, **83,** 719–726.

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, Hayward, CA: Institute of Mathematical Statistics.

Prentice, R.L. (1982), Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model, *Biometrika*, **69,** 331–342.

Prentice, R.L. and Breslow, N.E. (1978), Retrospective Studies and Failure Time Models, *Biometrika*, **65,** 153–158.

Ratcliffe, S.J., Guo, W. and Ten Have, T.R. (2004), Joint Modeling of Longitudinal and Survival Data via a Common Frailty, *Biometrics*, **60,** 892–899.

Ribaudo, H.J., Thompson, S.G., and Allen-Mersh, T.G. (2000), A Joint Analysis of Quality of Life and Survival Using a Random Effect Selection Model, *Statistics in Medicine*, **19,** 3237–3250.

Ringash, J., Bezjak, A., O'Sullivan, B., and Redelmeier, D.A. (2004), Interpreting Differences in Quality of Life: The FACT-H&N in Laryngeal Cancer Patients, *Quality of Life Research*, **13,** 725–733.

Ripatti, S. and Palmgren, J. (2000), Estimation of multivariate frailty models using penalized partial likelihood, *Biometrics*, **56,** 1016–1022.

Rizopoulos, D., Verbeke, G., Lesaffre, E., and Vanrenterghem, Y. (2008), A Two-Part Joint Model for the Analysis of Survival and Longitudinal Binary Data with Excess Zeros, *Biometrics*, **64,** 611–619.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008), Shared Parameter Models under Random Effects Misspecification, *Biometrika*, **95,** 63–74.

Robinson, G.K. (1991), That BLUP is a Good Thing: The Estimation of Random Effects, *Statistical Science*, **6,** 15–51.

Rothman, K.J. (2002), *Epidemiology: An Introduction*, Oxford University Press.

Schall, R. (1991), Estimation in Generalized Linear Models with Random Effects, *Biometrika*, **78,** 719–727.

Sen, P.K. and Singer, J.M. (1993), *Large Sample Methods in Statistics*, Chapman & Hall, New York.

Solomon, P.J. and Cox, D.R. (1992), Nonlinear Components of Variance Models, *Biometrika*, **79,** 1–11.

Song, X., Davidian, M., and Tsiatis, A.A. (2002), "A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data, *Biometrics*, **58,** 742–753.

Song, X. and Wang, C.Y. (2007), Semiparametric Approaches for Joint Modeling of Longitudinal and Survival Data with Time-Varying Coefficients, *Biometrics*, **64,** 557–566.

Stefanski, L.A. and Carroll, R.J. (1987), Conditional Scores and Optimal Scores for Generalized Linear Measurement- Error Models, *Biometrika*, **74,** 703–716.

Stiratelli, R., Laird, N., and Ware, J.H. (1984), Random Effects Models for Serial Observations with Binary Responses, *Biometrics*, **40,** 961–971.

Terrell, J.E., Ronis, D.L., Fowler, K.E., Bradford, C.R., Chepeha, D.B., Prince, M.E., Teknos, T.N., Wolf, G.T., and Duffy, S.A. (2004), Clinical predictors of quality of life in patients with head and neck cancer, *Archives of Otolaryngology Head & Neck Surgery*, **130,** 401–408.

Therneau, T.M., and Grambsch, P.M. (2001), *Modeling Survival Data*, New York: Springer.

Thisted, R.A. (1988), *Elements of Statistical Computing*, Chapman & Hall.

Tierney, L. and Kadane, J.B. (1986), Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of American Statistical Association*, **81,** 82–86.

Troxel, A.B. and Esserman, D.A. (2004), Frailty Models for Quality of Life in Oncology, *Journal of Biopharmaceutical Statistics*, **14,** 145-154.

Tseng, Y.K., Hsieh, R., and Wang, J.L. (2005), Joint Modelling of Accelerated Failure Time and Longitudinal Data, *Biometrika*, **92,** 587–603.

Tsiatis, A.A. (1981), A Large Sample Study of Cox's Regression Model, *The Annals of*

Statistics, **9,** 93–108.

Tsiatis A.A., Degruttola, V., and Wulfsohn M. (1995), Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS, *Journal of American Statistical Association*, **90,** 27–37.

Tsiatis, A.A. and Davidian M. (2001), A Semiparametric Estimator for the Proportional Hazards Model with Longitudinal Covariates Measured with Error, *Biometrika*, **88,** 447–458.

van der Vaart, A.W. (1998), *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A.W. and Wellner, J.A. (1996), *Weak Convergence and Empirical Processes*. New York: Springer–Verlag.

Verbeke, G. and Lesaffre, E. (1996), A Linear Mixed-effects Model with Heterogeneity in the Random-effects Model with Heterogeneity in the Random-effects Population, *Journal of American Statistical Association*, **91,** 217–221.

Verbeke, G. and Molengerghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, Springer-Verlag, New-York.

Waclawiw, M.A. and Liang, K.Y. (1993), Prediction of Random Effects in the Generalized Linear Model, *Journal of American Statistical Association*, **88,** 171–178.

Wang, Y. and Taylor, J.M.G. (2001), Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome, *Journal of American Statistical Association*, **96,** 895–905.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989), Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions, *Journal of American Statistical Association*, **84,** 1065–1073.

Wu, M. and Bailey, K. (1989), Estimation and Comparison of Changes in the Presence of Informative Right Censoring: Conditional Linear Model, *Biometrics*, **45,** 939–955.

Wu, M. and Carroll, R. (1988), Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modelling the Censoring Process, *Biometrics*,

44, 175–188.

Wulfsohn M. and Tsiatis A.A. (1997), A Joint Model for Survival and Longitudinal Data Measured with Error, *Biometrics*, **53,** 330–339.

Xu, J. and Zeger S. (2001a), The Evaluation of Multiple Surrogate Endpoints, *Biometrics*, **57,** 81–87.

———(2001b), "Joint Analysis of Longitudinal Data Comprising Repeated Measures and Times to Events, *Applied Statistics*, **50,** 375–387.

Zeger, S.L. and Karim, M.R. (1991), Generalized Linear Models with Random Effects: a Gibbs Sampling Approach, *Journal of American Statistical Association*, **86,** 79–86.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988), Models for Longitudinal Data: A Generalized Estimating Equation Approach, *Biometrics*, **44,** 1049–1060.

Zeng, D. and Cai, J. (2005a), Simultaneous Modelling of Survival and Longitudinal Data with an Application to Repeated Quality of Life Measures, *Lifetime Data Analysis*, **11,** 151–174.

———(2005b), Asymptotic Results for Maximum Likelihood Estimators in Joint Analysis of Repeated Measurements and Survival Time, *The Annals of Statistics*, **33,** 2132–2163.

Zhang, D. and Davidian, M. (2001), Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data, *Biometrics*, **57,** 795–802.

Ye, W., Lin, X.H., Taylor, and J.M.G. (2008), A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data, *Statistics and Its Interface*, **1,** 33–45.